Project Report

Peer Review of the MISTI Survey and Evaluation Methodology

RAND NDRI

PR-1334-MSI
September, 2014
Prepared for Management Systems International (MSI)

NOT CLEARED FOR PUBLIC RELEASE

Preface

This research was conducted within the International Security and Defense Policy Center of the RAND National Security Research Division (NSRD). NSRD conducts research and analysis on defense and national security topics for the U.S. and allied defense, foreign policy, homeland security, and intelligence communities and foundations and other nongovernmental organizations that support defense and national security analysis.

For more information on the International Security and Defense Policy Center, see http://www.rand.org/nsrd/ndri/centers/isdp.html or contact the director (contact information is provided on the web page).

Comments or questions on this draft report should be addressed to the project leader,

Table of Contents

Figures

Tables

Abbreviations

IDLG	Independent Directorate of Local Governance
MISTI	Measuring Impact of Stabilization Initiative
MRRD	Ministry of Rural Rehabilitation and Development
NSP	National Solidarity Program
SIKA	Stabilization in Key Areas
USAID	United States Agency for International Development

Summary

This review assesses the methodology and tools developed by the Measuring Impact of Stabilization Initiative (MISTI) project to evaluate the impacts of United States Agency for International Development (USAID) stabilization programming in Afghanistan. Specifically, we were asked to assess the quasi-experimental impact evaluation methods designed to evaluate the impacts of this development programming on Afghans' perceptions of stability.

Seven key challenges were identified during our reviews of MISTI data and existing reports. The first two challenges stem from difficulties faced by the evaluation team in coordinating with the implementing partners which, in turn, led to problems in (1) identifying intervention villages and (2) understanding the implementing partners' theories of change. The third and fourth are data-related challenges, namely (3) a lack of comprehensive historical data on development programming and (4) a lack of a credible metric for measuring support for the Taliban as compared to the Afghan government. Challenges 5 and 6 relate to technical implementation, specifically the difficulties faced in (5) identifying appropriate control villages and (6) developing a defensible metric of stability. The final challenge (7) was a design challenge in that the villages benefiting from the stability programming may not be necessarily representative of the overall population, threatening the external validity of the MISTI evaluation for other stability-focused programming in Afghanistan, or elsewhere in the world.

Several key overall findings emerge from our review of the MISTI approach. First, the MISTI data collection effort likely provides an effective tool for measuring the *direct* impacts of USAID programming conducted during 2012-2015, meaning the impacts for which these programs were originally designed. That is to say, MISTI should be able to provide credible estimates of governance programs on governance outcomes, economic programs on economic outcomes, etc. Although the evaluation approach presented in existing MISTI reports has several empirical limitations, most are surmountable given the range and quality of data collected, making a credible impact evaluation of direct program impacts feasible.

Second, though the MISTI evaluation is well-designed to measure these direct effects of USAID programming, it is less clear whether MISTI will be able to provide credible estimates of the impact on perceptions of stability. Although these outcomes have been the focus of the MISTI evaluation, they are essentially indirect potential outcomes of the various programs put into place. The evaluation, however, lacks a clearly delineated theory of change for the each stabilization program to explain how the programs could influence perceptions of stability. Without such a framework it would be difficult to interpret a positive result even if that was the result obtained. Further, the primary tools used to measure the perceptions of stability – the stability index and endorsement experiment – are unlikely to be well suited for measuring either stability or relative support for the Taliban. The stability index in particular is poorly defined,

combining fairly disparate elements which do not add up to a clear construct for 'stability'; importantly, it is unlikely that a clear 'stability' construct exists or is meaningful for this type of impact evaluation. And the available data from the endorsement experiment suggests that the approach was unable to capture individuals' support for non-state actors. With regard to stability, we suggest that there is still scope for analyzing this important outcome, for example by considering component parts if the index that are well defined.

Third, a key lesson learned from this evaluation—and a more general one – is the importance of external coordination in planning and carrying out an impact evaluation in a developing, unstable country context like Afghanistan. This would involve, most importantly, close coordination with both the implementing partners and other development, security, and international organizations. This coordination would mean that evaluators and implementers are in communication to ensure, for example, that unambiguous data on program location and the theories of change underlying the efforts of the implementing partners are available to evaluators. Doing so involves additional time, effort, and resources for the implementing partners, but the cost is reasonable relative to the benefits; as this review has shown, the lack of such coordination has been particularly problematic for the MISTI evaluation.

Our review also produces the following recommendations for improving future iterations of MISTI analysis as well as future MISTI-like efforts:

- <u>Recommendations for Improving MISTI Waves 4 and 5 Reports</u>
  1. Assess severity of treatment/control misspecification by augmenting existing MISTI validation effort with a village- or district-level survey module during Wave 5 data collection.
  2. Conduct power calculations in order to assess whether the MISTI data have sufficient program villages to measure program effect.
  3. Use propensity score-based quasi-experimental methods (e.g., IPT, CBPS) in addition to only "exact matching" methods and indicate whether the findings are robust to choice of method.
  4. Work with implementing partners to identify how villages were selected for program participation.
  5. Include expanded data and project-specific variables on development programming.
  6. Use data-driven methods for deriving the requisite stability index.
  7. Analyze individual or groups of components of stability separately for the impact evaluation.
  8. Validate the stability measure using data from 2012-2013
  9. Coordinate with ISAF and other representatives to validate the stability metric
  10. Rather than focus on only the reduced form outcomes currently considered—i.e., from program inputs to stability – the analysis should also evaluate whether the

program is having the intended immediate impact (e.g., improved district governance) as well.

- Things that Could Have Been Done and Should be Done for Future Evaluations:
  1. Coordinate with implementing partners from the onset in developing and implementing the impact evaluation, in particular to be able to clearly articulate identify program areas.
  2. Conduct power calculations before commencement of the evaluation or, if power calculations were done in advance of data collection, discuss the power calculations used in deciding on the data collection plan.
  3. Gather comprehensive information on previous development programming and use this as controls in the analysis or as a basis for stratification in the analysis.
  4. Cleary articulate a theory of change at program commencement.

1. Intent of This Review

The Measuring Impact of Stabilization Initiative (MISTI) was contracted by the United States Agency for International Development (USAID) to provide third-party monitoring of its programming in Afghanistan.[1] MISTI was requested to provide two types of analytic support. The primary purpose was to develop a methodology for conducting impact evaluations of USAID programming being conducted by the USAID Stability Unit. The secondary purpose was to develop tools for assessing long-term stability trends in districts with USAID programming.

The review examines the methodology and tools developed by MISTI both to measure stability trends and to evaluate impacts. The intent of this review is to identify strengths and limitations of the approach, and identify recommendations for improvement. For this review, the research team analyzed MISTI reports, surveyed relevant academic research, and conducted interviews with relevant USAID program officers, MSI Survey Specialists, and MISTI researchers.

We have identified and focus on seven key challenges that the MISTI impact evaluation faced. These challenges are (1) identification of USAID program villages, (2) the search for appropriate controls for the impact evaluation, (3) adjusting for previous development programming, (4) the design of an appropriate stability index, (5) measuring support for the Taliban as compared to the Afghan government, (6) articulating an appropriate theory of change, and (7) ensuring external validity of the evaluation. For each of these challenges we outline the issue and then discuss MISTI's approach for addressing the challenge, any remaining issues and how they might be resolved by the research team, and conclude by identifying implications for evaluation research in conflict settings.

The report begins by providing general background information on the USAID programs being evaluated, the household surveys being collected, and the overall structure of MISTI's impact evaluation and stability trend tracking design. The subsequent seven sections discuss each of the challenges just listed. A final section draws out the implications of these challenges for, respectively, the analysis of stability trends and the evaluation of the impact of this USAID programming.

[1] This work was funded under AID-306-TO-12-00004.

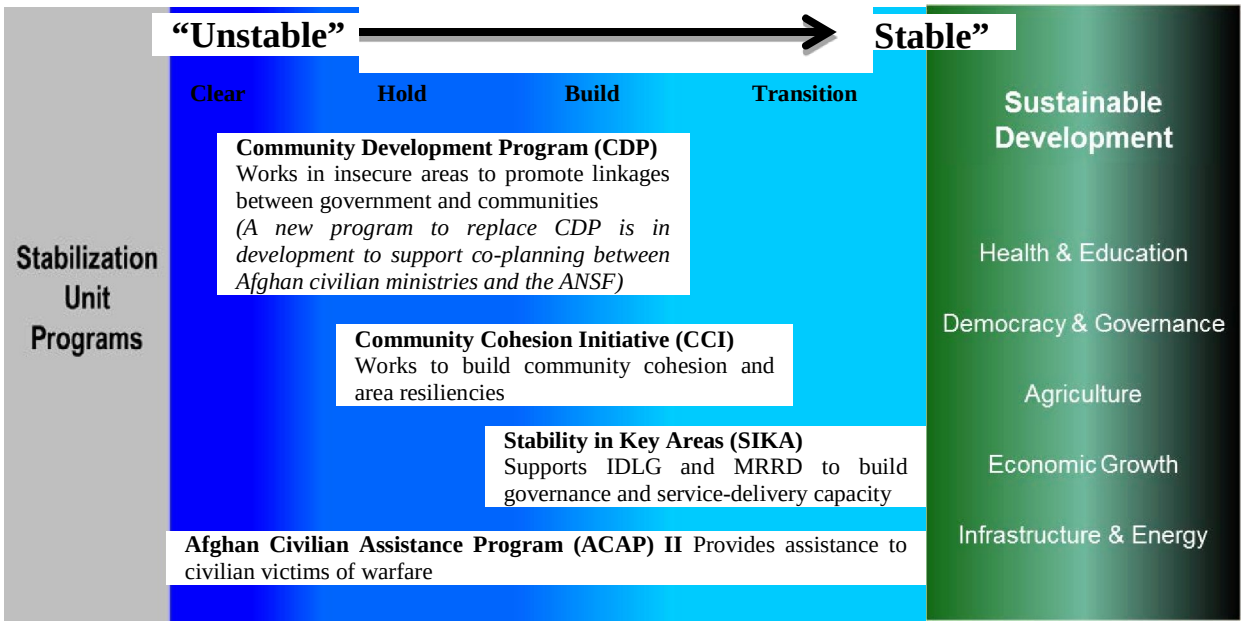
2. Background on MISTI Evaluation

Programs Being Evaluated

The MISTI impact evaluation is focused on programming implemented through the USAID Stabilization Unit during 2012-2015. This programming is directed at "reinforcing the legitimacy of the [Afghan] government and its effectiveness at the subnational and community levels, and improving communities' resilience to malign, antigovernment actors".[2] Thus, it differed from earlier Stabilization Unit programming (2010-2011) which was focused on supporting counterinsurgency efforts in Afghanistan.

The suite of Stabilization Unit programming is described in Figure 1. The MISTI impact evaluation focuses on the Community Cohesion Initiative and the Stabilization in Key Areas programs. The evaluation was also asked to include the Kandahar Food Zone program, as described below.

Figure 1: USAID Stabilization Unit Programming during 2012-2015



Note: From "USAID Stabilization Unit Afghanistan: Performance Management Plan, Fiscal Years 2012-2105".

The Stabilization in Key Areas (SIKA) program is intended to promote stability by supporting Afghan government efforts to implement community-led development and governance initiatives. This program is implemented in coordination with the Ministry of Rural Rehabilitation and Development (MRRD) and Independent Directorate of Local Governance

[2] "USAID Stabilization Unit Afghanistan: Performance Management Plan, Fiscal Years 2012-2105" (p. 2).

(IDLG).[3] Importantly, this program is implemented through four separate offices (and task orders) and involves a diversity of international and local partners:

- SIKA-East: Managed by AECOM but implemented through (1) International Relief and Development, (2) Development Transformation, (3) GardaWorld, (4) Technologists Inc., and (5) Overseas Strategic Consulting, LTD;[4]
- SIKA-North: Managed by DAI but implemented through (1) ACSOR Surveys, (2) Pax Mondial, (3) Sayara, (4) The Liaison Office, (5) Training Resources Group, and URS Corporation;[5]
- SIKA-South: Managed by AECOM but implemented through (1) International Relief and Development (IRD), (2) Technologists Inc., and (3) Overseas Strategic Consulting, LTD;[6] and
- SIKA-West: Managed by AECOM but implemented through (1) International City/County Management Association (ICMA), (2) GardaWorld, (3) Technologists Inc., and (4) Overseas Strategic Consulting, LTD.[7]

Although this is beyond the scope of the current review, which is focused on assessing the impact evaluation itself, this multitude of international actors working through a single national-level mechanism is reported by personnel involved in the MISTI impact evaluation to have created coordination problems for the MRRD and IDLG personnel working with SIKA.[8]

The Community Cohesion Initiative (CCI) programs are intended to enhance resilience by strengthening connections across communities and between communities and the Afghan government.[9] CCI is administered through the provision of clusters of small grants at the district- or village-level. CCI was implemented through two offices with Creative working in the East of Afghanistan and International Organization for Migration working in the North and West.[10]

---

[3] "USAID Stabilization Unit Afghanistan: Performance Management Plan, Fiscal Years 2012-2105" (p. 6).

[4] AID-306-C-12-00002 (p. 46). SIKA-East is a $177 million dollar effort.

[5] AID-306-C-12-00003 (p.44). SIKA-North is a $46 million dollar effort.

[6] AID-306-C-13-00003 (p.50). SIKA-South is a $60 million dollar effort.

[7] AID-306-C-12-00004 (p.41). SIKA-West is a $63 million dollar effort.

[8] As this could have important implications for the effectiveness of the programs themselves, a separate qualitative evaluation effort exploring this issue may be warranted. Note that, in addition to this, at least one of the implementing teams faced challenges in the rapid turnover of personnel ("Stability in Key Areas – West: Mid-term Performance Evaluation", 26 March 2014).

[9] "USAID Stabilization Unit Afghanistan: Performance Management Plan, Fiscal Years 2012-2105" (p. 6).

[10] We were unable to track down the relevant task orders for the CCI projects.

Lastly, the Kandahar Food Zone (KFZ) is designed to discourage poppy cultivation, and encourage "licit economic growth", and enhance the perceived effectiveness and legitimacy of Afghan government institutions.[11]

Table 1 provides examples of specific projects being implemented by each of the seven programs included in the evaluation. Each program includes both infrastructure and vocational training projects, while a few of the programs also include community cohesion, outreach, or other similar programming.

Table 1: Types of Projects

USAID Program		Example Projects†
SIKA	East	School Boundary Wall Constrcution
		Embroidery Vocational Course Training & Embroidery Machine Supply
		Tertiary Road Gravelling
		Digging new wells and installing hand pumps
	North	Increasing Stability among Youth through a University Prep Exam Course (Concur)
		Connecting GIRoA with farmers through celebrating New Year and farmer's day
		Linking Afghan Youth with District Entities through English Courses
		Increasing Community's Trust on the District Government through village road gravelling
	South	Workshop: Access to GIRoA Services
		Irrigation Project
		Road and/or Culverts
		Vocational Training
	West	Mobile Phone Repair Vocational Training
		Culvert Construction
		Community Based Health Education Course (CHA)
		Tailoring Vocational Training Program
CCI	Creative	Tailoring Training: Income Generation Opportunities for Female Youth
		Promoting Justice and Stability through Conflict Resolution Training
		Extending the reach of peaceful elections messaging through radio
		Sahib Jan Rural Sub-Road Repair: GIRoA Connecting Communities
	IOM	Guzara Youth Art Peace Championship: Promoting Peace and Tolerance
		Adraskan Football Tournament: Sports for Tolerance
		Promoting peaceful transition on Highway 1: Equipping the Jahan Khan School
		Promoting Stability Around the Airport: Rehabilitating Retaining Wall, Gawashan
KFZ		Women's Vegetable Production and Processing Project
		Vocational Training in Mechanical Repairs of Machinery
		Check Dam Construction Asodah
		Irrigation Canal Sangi-Hisar

†: Example projects are designed to be representative of the range of different types of projects implemented by each implementing partner. However, they were not selected to be proportionally representative.

[11] USAID Cooperative Agreement AID-306-A-13-00008.

MISTI Data Collection Effort

Household survey data for MISTI is collected semi-annually, with the first wave collected in Fall 2012. As a result of the large sample size of each wave – 35-40,000 households, depending on the wave – data collection for each of the first three waves of available data has taken 3-4 months. Data from wave 4 of the survey is now being collected and prepared for analysis.

The sample sizes (number of villages and households) in each wave and the number of treatment villages in each are presented in Table 2. The "Cross section data" in rows 1-3 show the total samples in each wave. Additionally, as the MISTI evaluation relies on the repeated observations of villages across waves, (the village level panel data) the table reports (in rows 4-6) these key summary statistics as well as the number of USAID treatment villages for each of these panels.

Table 2: MISTI Household Data Key Summary Statistics

		# of Districts	# of Villages	# of Households (per wave)
Cross-sectional data	Wave 1	83	2,228	34,972
	Wave 2	82	2,373	36,475
	Wave 3	93	2,559	40,405
Village-Level Panels	Waves 1 & 2	64	890	13,709
	Waves 1 & 3	63	979	15,394
	Waves 2 & 3	75	1,598	24,627
	Waves 1, 2 & 3	60	706	10,923

Note: The first three rows report data on the complete sample. The bottom four rows report number on the number of households for the wave with the minimum number of households as the power is relative to this minimum number.

MISTI Impact Evaluation

The first intent of the MISTI evaluation is to measure the impact of USAID programming on (1) stability and (2) support among the population for the Afghan government. Data on these two key outcomes are collected through household level surveys in districts where USAID programming is focused. Thus, the evaluation focuses on assessing the impact of USAID programming on households' perceptions of stability and reported support for the Afghan government.

The MISTI evaluation is focused at the village level, as USAID projects are implemented at that level. Essentially, it compares whether and how households in villages with USAID programs differ from those without USAID programming. How this comparison is made is the key to the validity of the results of the evaluation.

Ideally, a randomized control trial (RCT) would be conducted in which projects are randomly allocated to a subset of villages from a pre-selected pool of eligible villages. Data would be

collected in villages that receive projects ("treatment" villages) and those that do not ("control" villages). However, an RCT was not feasible in the Afghan context.

Instead, the MISTI evaluation combines two standard alternative "quasi-experimental" approaches from the evaluation toolbox: (1) matching and (2) "difference-in-difference". The matching approach tries to mimic a random program allocation by comparing villages with USAID projects to other villages that are very similar. In the MISTI evaluation, this matching is done following data collection, so that project villages in the household survey are combined to non-project villages in the household survey with similar characteristics.[12]

The difference-in-difference approach similarly helps adjust for possible non-randomness in project allocation. Specifically, in the MISTI context, it is likely that pre-existing (i.e., before USAID intervention) levels of stability and government support influenced village selection. Thus, a comparison of these outcomes after the stability programs were implemented may merely reflect differences that existed before the programs. Rather than compare levels of the outcomes across program and non-program villages, the difference-in-difference approach compares the changes in outcomes in these two groups. Thus, data is collected from villages both before the program begins ("baseline") and following implementation ("follow-up"). By comparing changes rather than levels, the approach controls for the initial or baseline differences in outcomes.

We should note that while both approaches have well known limitations, especially relative to a randomized trial, they are valid under certain assumptions and are widely used by researchers. Our focus in this review therefore is not the general validity of the approaches but rather in their application to the present context, and specifically whether data constraints or other factors may affect the reliability of the MISTI findings.

Stability Trends

The second intent of the MISTI program, in addition to impact evaluation, is to trace stability trends over time in the districts where USAID programming is being implemented. The approach that MISTI uses for tracking these is a district-level stability index, which is compared over time and to districts without USAID programming. This allows MISTI to "inform USAID decision makers and implementing partner managers of changes in stability occurring in the districts where USAID stabilization programming is taking place across Afghanistan, and control districts, and help identify improvements and declines in stabilization in their areas of responsibility".[13]

Stability is measured using two approaches. The first approach explores, individually, ten separate dimensions of stability based on the MISTI household survey. These categories are: (1)

[12] Matching can also be done ex ante if sufficient secondary data are available.

[13] "MISTI Stabilization Trends and Impact Evaluation Survey Analytical Report, Wave 3" (Preliminary Draft).

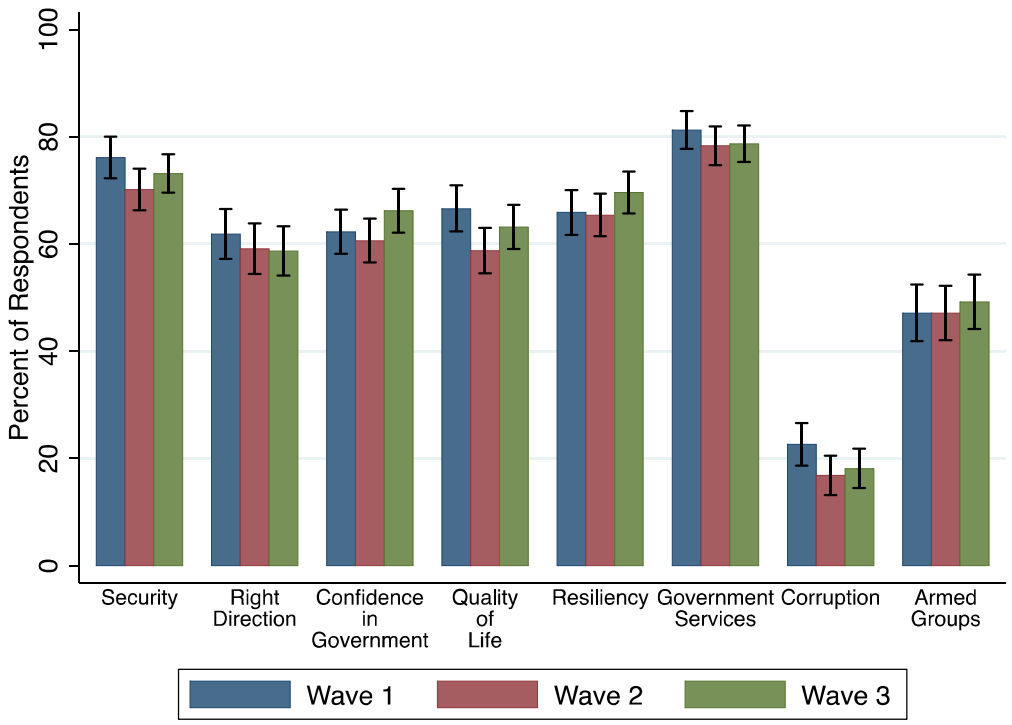
security and crime, (2) governance, (3) service provision, (4) rule of law, (5) corruption, (6) quality of life, (7) economic activity, (8) community cohesion, (9) grievances, and (10) media. MISTI reports potential trends in each of these.

The second, which is the focus of this review and most policy discussions surrounding MISTI, is a stability index. This index includes: (1) 35 variables from the household survey, which are aligned along 8 different dimensions (illustrated in Figure 2); (2) qualitative assessments from survey enumerators of local conditions; and (3) an overall assessment of district permissiveness; and (4) the number of security incidents.[14] It will be noted that these measures combine subjective perceptions of household respondents and those of enumerators in addition to objective measures such as the number of security incidents. Importantly, the stability index is also used for the impact evaluation, which is discussed further in the subsequent section, "Challenge #4: Design of Stability Index". For now, to put this review in perspective, we present some general illustrative trends in stability measures as well as how they are correlated.

Figure 2 provides general summary statistics for 8 of the 35 variables selected from the household survey; one variable from each of the eight different dimensions was selected for illustrative purposes. These summary statistics include data both from villages with USAID programming and those without, and districts with USAID programming and those without, so the trends should not be interpreted as showing possible impacts of the program. Rather they are presented to illustrate how these characteristics have evolved through the first three waves of data collection.

[14] The weighting of these four types of variables is 75%, 10%, 10%, and 5%.

Figure 2: Means for Key Stability Measures



Note: Figure reports average response to eight key variables used in construction of MISTI stability index. Analysis is restricted to only those villages that were included in all three waves of data. Standard errors reflect clustering at the village level and reported means for each wave reflects controlling for potential changes in the composition of the sample along three dimensions: gender, income, and age. Bars report 95% confidence intervals of estimation.

The pairwise correlations of these eight variables in reported in Table 3; all reported correlations, with one exception, are significant within the data at the 1% significance level (not accounting for possible sampling error). Several of these measures are strongly correlated – e.g., overall feeling of security is strongly positively correlated with (1) belief that Afghanistan was moving in the right direction, (2) confidence in government, (3) quality of life, (4) view toward government service provision, and (5) absence of armed groups. However, several of these measures are only weakly correlated. The implications of this heterogeneity for the MISTI evaluation findings using the stability metric are discussed later in Chapter 6.

Table 3: Pairwise Correlation of Components of Stability Index

	Security	Right-Direction	Confidence-in-Government	Quality-of-Life	Resiliency	Government-Services	Corruption	Armed-Groups
Security	1							
Right-Direction	0.44	1						
Confidence-in-Government	0.28	0.3	1					
Quality-of-Life	0.31	0.36	0.25	1				
Resiliency	0.07	0.1	0.08	0.1	1			
Government-Services	0.33	0.31	0.23	0.27	0.05	1		
Corruption	0.03	0.03	0.01	0.04	-0.02	0.06	1	
Armed-Groups	0.18	0.17	0.12	0.11	0.02	0.09	0.05	1

Note: All reported correlations are significant at the 1% level with the exception of the correlation between corruption and confidence in government; this significance does not account for sampling error.

Assessing Support for the Taliban vs. Afghan Government

A key outcome variable in the impact evaluation analysis is support for the Taliban relative to the Afghan government among local nationals. Specifically, the impact evaluation wants to test whether the USAID programming is increasing support for the Afghan government relative to support for the Taliban.

The impact evaluation uses the household survey to conduct an endorsement experiment to assess the relative support for these two national actors. The "experiment" is relatively straightforward. In each village 50% of the respondents are assigned to the "Taliban" group and 50% are assigned to the "Afghan government" group. In each case the respondents are asked a series of questions about political issues, with the question indicating that the assigned group (e.g., Taliban vs. Afghan government) supported that group:

> *"Sample Question: It has recently been suggested by the Afghan government [Taliban] that people be allowed to vote in elections to select the members of their district council. Do you oppose or support such a policy, or are you indifferent to this policy? Do you strongly or only somewhat oppose/support?"*

This approach allows the evaluation to estimate support for the Taliban as compared to the Afghan government by comparing the two groups. Specifically, if villages are more likely to

respond favorably to questions associated with the Taliban, then that would be used as an overall assessment of sympathy for the Taliban in those villages.

3. Challenge #1: Identification of USAID Program Villages

Description of the Challenge

Difficulties identifying which villages in the MISTI household survey received USAID programming will make it difficult for the MISTI evaluation to measure program effects. The issue is one of misclassification of treatment and control status – either (1) program villages are misclassified as control villages or (2) control villages are misclassified as treatment villages, or both.[15] Misclassification is a form of measurement error in a variable. If this error is random, it will, all things equal, lead to estimates that are biased toward zero – i.e., toward finding no effect.[16] Importantly, even a small amount of misclassification (e.g., under 5%) could make program effect impossible to detect.[17]

A related problem is that the existing data may not have a sufficient number of treatment villages to allow MISTI to measure program effects. Though sample sizes of approximately 100 treatment villages ("clusters") and 100 control villages are standard in most empirical work, power calculations can be used to calculate the number of villages required to detect program effect.

The Challenge in the MISTI Context

The MISTI impact evaluation team faced, and continues to face, significant problems in identifying the exact villages in which USAID programming is implemented and which it is not. The implementing partners (IPs) were not prepared to support an impact evaluation, which created two challenges. First, the IPs had not agreed, either amongst themselves or with the MISTI impact evaluation team, on a common village sampling frame for Afghanistan, from which villages would be selected. Second, once the selections of villages were made using the sampling frames at hand, detailed information on the specific villages getting the programs was not systematically recorded by the IPs.

The data provided to MISTI by the IPs, therefore, had several shortcomings that complicated accurate identification of USAID program villages. These challenges were identified by the

[15] Though classical measurement error that is uncorrelated with the unobservables will simply cause attenuation of the point estimates (e.g., Black et al., 2003), measurement error in binary regressions is typically not classical because errors are mean-reverting (e.g., Aigner 1973, Kreider 2010). Millimet (2010) provides a review of the general challenges facing empirical analyses with measurement error in a binary regressor.

[16] If misclassification is severe – in that more than half of villages in the sample are misclassified – sign reversal is possible (e.g., Frazis and Lowenstein 2003)—that is, the estimate could be negative when the true program effect is positive.

[17] E.g., Kreider (2010).

MISTI staff. First, the specific geospatial data (i.e., latitude and longitude) provided for villages benefiting from the USAID programming often did not match with other data provided for that same village (e.g., the locational information for a project implied it was in district X while the data provided by the implementing partner indicated that it was in district Y). Second, locational data provided for projects was often incongruous with the type of project being implemented (e.g., a culvert project cannot be implemented in a desert), hence clearly inaccurate. Third, certain classes of USAID programs by their nature affect more than a single village (e.g., roads), but only one village location was provided. Fourth, the locational information provided to MISTI did not always reference a populated area.

Table 4 provides an indication of the extent of this problem, though we are unable to precisely assess the extent.[18] The table includes a summary of (1) implementing partner data provided to the MISTI impact evaluation team and (2) the projects that for which the MISTI impact evaluation team was able to verify location information. The data for SIKA-West provide a clear example of what could be the first type of classification error (programming villages misclassified as control villages) as the MISTI project tracker only includes 123 projects in 93 villages while the implementing partner provides data on 257 projects across 246 locations. Conversely, the data for SIKA-North provides a suggestive example of the second type of classification challenge as the MISTI project tracker reports data on more locations that the implementing partner.

Table 4: Summary Statistics for Potential Project Misclassification

		Implementing Partner Data		MISTI "USAID Project Tracker"		
		# Projects	# Villages/ Implementation Points	Projects	Total # of Villages	# Villages in Waves 1, 2, or 3 of MISTI
SIKA	East	230*	209*	249	201	109
	North	534	172	508	192	97
	South	413	2398§	249	539	197
	West	257•	246*	123	93	68
CCI	Creative	725	475	†	†	†
	IOM	135	114	†	†	†
KFZ		50	‡	†	†	†

†: No data in project tracker provided JUL 2014.
‡: No data provided by implementing partners.
*: Includes completed, on-going, and closed projects only.
§: This does not use the unique locational information but rather the raw counts of the number of locational points associated with each project.

[18] The MISTI impact evaluation team is currently working to remedy this issue, so the severity of this issue may be reduced for subsequent analyses.

The data from Table 4 also provide information on the potential secondary problem discussed at the beginning of this section – i.e., whether there are sufficient treatment villages in the sample to measure program effect. Though there are only 108 total treatment villages in the overall panel, this suggests that the data from Waves 4 and 5, if appropriately collected, could allow a panel with nearly 500 treatment villages (109 + 97 + 197 + 68 = 471) which would almost certainly be sufficient for this impact evaluation. Specifically, as 471 known treatment villages were included in at least one of the first three waves, there could be data for at least two points in time (which would allow difference-in-difference estimation) for each of these villages if they are included in the data collection from either Wave 4 or 5.

MISTI Remedy and Remaining Issues

The MISTI evaluation team implemented several remedial measures to address these known challenges.[19] This included project-by-project verification of locations using overhead satellite imagery, additional communication with IPs about project location(s), and geo-located photographs of projects being implemented. This verification process expanded the pool of verified USAID programming villages that could be included in the analysis.

However, despite these remedial efforts, there is significant uncertainty about the full range of completed, ongoing, and planned projects and the locations in which those projects were implemented. Thus, there are likely to be significant classification of the first type – i.e., villages labeled as "controls" in the data where programs *are* being implemented.

The other remaining issue, discussed above, is that power calculations should be conducted to estimate the minimum number of treatment villages requisite for estimating program effect. Importantly, if the intent is to stratify the analysis by program, then separate power calculations should be done for each program.

Implications for Impact Evaluations in Conflict/Complex Settings

A best practice in the impact evaluation literature is close coordination between the IPs and the evaluation team before both the evaluation and implementation begin.[20] In a setting like Afghanistan, without standardized data frames, this coordination is of particular importance and perhaps the only way to address this issue. Coordination must include both the use of a common

[19] These remedial measures were implemented under a separate USAID task order.

[20] As an example, see Jones et al. (2009). And USAID (2013) reports that: "Impact evaluations are always most effective when planned before implementation begins. Evaluators need time prior to implementation to identify appropriate indicators, identify a comparison group, and set baseline values. **In most cases they must coordinate the selection of a treatment and comparison group with the implementing partners.** If impact evaluations are not planned prior to implementation the number of potential evaluation design options is reduced, often leaving alternatives that are either more complicated or less rigorous. As a result, Missions should consider the feasibility of and need for an impact evaluation prior to and during project design." (bold faced text added by author)

frame, and once villages are selected, clear recording of location and other identification information for program villages.

A related best practice is conducting power calculations when the evaluation is first being designed. Power calculations allow researcher to be judicious in data collection by limiting evaluation to the number of villages needed to assess program effect.

4. Challenge #2: Finding Appropriate Control Villages

Scope of the Problem

The "gold standard" in impact evaluation is randomized program assignment, using baseline and follow-up surveys. In the MISTI context, this approach would require that the IPs (1) identify a sample of villages that qualify for the program and (2) randomly select some villages to receive programs and others to receive nothing. Alternatively and usually more feasibly, the latter group would also receive the programs, but only after an interval, during which they can still serve as controls. However, as randomization was not feasible, "quasi-experimental" methods must be used.[21]

The intuition of quasi-experimental methods is relatively straightforward. A key characteristic of randomization is that the only difference, statistically speaking, between "treatment" and "control" villages is that some received the intervention and others did not; treatment and control villages should therefore be identical on average (that is, allowing for normal variation across villages). Quasi-experimental methods such as matching try to mimic this equivalence by using available data from treatment and control villages to find a group of control villages that is as similar as possible to the villages that received the program, based on data on characteristics of the villages.

A variety of matching methods are now available to the empirical researcher. Regardless of the specific empirical approach used, the set-up must meet the same assumption – i.e., the characteristics used to match communities should fully account for differences between the treated and control communities, so that the observed differences in that outcome can be attributed to the program being evaluated. This property is known as "unconfoundedness", or adequate "selection on observables".

There are at least five different empirical matching estimation methods that can be used in this context (e.g., Imbens 2004). However, approaches that rely on either (1) matching on the propensity score or (2) matching on covariates have been among the most popular in the impact evaluation literature.[22] The first rely on matching on a single variable that is a composite of many covariates of interest and the second rely on matching on many variables simultaneously.

Among the first type of estimators – i.e., those that rely on "matching on the propensity score" – the inverse probability weighting, or "IPW", (e.g., Wooldridge 2010) is the most

[21] Even if this randomization had been attempted, it is anticipated that there may have been significant challenges including non-compliance, other development donors targeting "control" villages, etc.

[22] This is largely a result of the prevalence of economists and political scientists within these fields. Bayesian methods (e.g., TWANG of Ridgeway et al. 2012) are popular among statisticians. Note that TWANG, as an example, is also easily implementable in many widely used statistical packages (http://cran.r-project.org/web/packages/twang/vignettes/twang.pdf).

intuitive and simplest to implement.[23] However, as misspecification of the propensity score model can lead to biased estimates, more recent propensity score methods have been developed that are either doubly robust – i.e., unbiased if either the propensity score model or regression model is correctly specified – (e.g., Inverse Probability Tilting ["IPT"] of Graham, Pinto, and Egel 2012) or robust to mild misspecification of the propensity score (e.g., Covariate Balancing Propensity Score ["CBPS"] of Imai and Ratkovic 2014).

The MISTI evaluation relies primarily on the second approach, i.e., matching on covariates. Such methods are attractive as finding control villages that have identical (or nearly identical) observable characteristics as treatment villages is an intuitive way to do matching; however, they do not have well-behaved statistical properties (e.g., Abadie and Imbens 2006) and in many cases, as is the case for MISTI discussed below, finding exact matches can be difficult.

The Challenge in the MISTI Context

The longest panel of data available for the impact evaluation, which includes data from Waves 1 and 3, contains a total of 113 treatment villages and 867 potential control villages.[24] The goal of the quasi-experimental methods is to identify control villages from the 867 that are similar to the 113 that benefited from USAID programming.

The effectiveness of these methods in producing reliable estimates for MISTI relies on the availability of sufficient data on the villages in both treatment and potential control villages. In particular, these methods assume that the researcher has access to data on all variables that affected the likelihood that a village received a project (or not).[25]

The MISTI impact evaluation uses two "exact matching" quasi-experimental approaches – "exact" meaning that they try to identify control villages that are the same as treatment village for all covariates believed to affect the likelihood of treatment assignment. The first uses an approach called coarsened exact matching (CEM) in which the observable data are "coarsened" (e.g., village size is coarsened from number of households to a discrete indicator for the villages have greater than or less than 1,000 individuals).[26] This process makes it easier to find matches than exact matching. The second, nearest-neighbor matching uses a Euclidean-distance like

[23] Inverse probability weighting is "easy" as it simply requires the researcher to (1) estimate a propensity score model in a first-stage and then (2) use the predicted values from this propensity score in a second-stage.

[24] This is based on the village-level data set "WAVE1-3 Data for Jay.xlsx" provides by the MISTI team. This includes all villages reported as treated by the third wave. Note that the "MISTI Wave 3" report indicates that only 108 villages are available for impact evaluation using these data (p.1).

[25] In addition, the propensity score model should include all other variables that will be used in the final analysis.

[26] E.g., Blackwell et al. (2009). A user-friendly overview of the program is provided at http://www.stata.com/meeting/boston10/boston10_blackwell.pdf.

intuition in terms of finding the control villages that are the "closest" across the observable characteristics.[27]

Although the current reports provide details on the approaches used, they do not clearly articulate whether either of these matching approaches is "successful" in identifying an appropriate set of control villages. Indeed, the only data provided (Table 2) compares the treatment villages and the full sample of potential control villages, not those selected as controls through the approaches that the impact evaluation team applied. In fact, the authors say in the footnote to Table 2 in this report that "This suggests that villages that have received assistance are different from 'average' villages in our survey, indicating that appropriate 'matches' for intervention villages may not be present in our survey sample" (p.162). However, the report does not provide sufficient data or further discussion regarding this assertion.[28]

We explore this potential challenge – i.e., that villages selected for USAID programming are perhaps too dissimilar to allow the use of matching methods – in Figure 3. For illustrative purposes, this figure uses a simple specification of the propensity score approach, a probit in which each covariate appears linearly, to explore the similarity of treatment and potential control villages.[29] Importantly, these distributions are qualitatively similar (as can be seen below) and a standard quantitative test for "balancing", which assesses whether there are sufficient control villages to construct a set of control villages that mimic randomization,[30] suggests that these distributions are sufficiently similar.

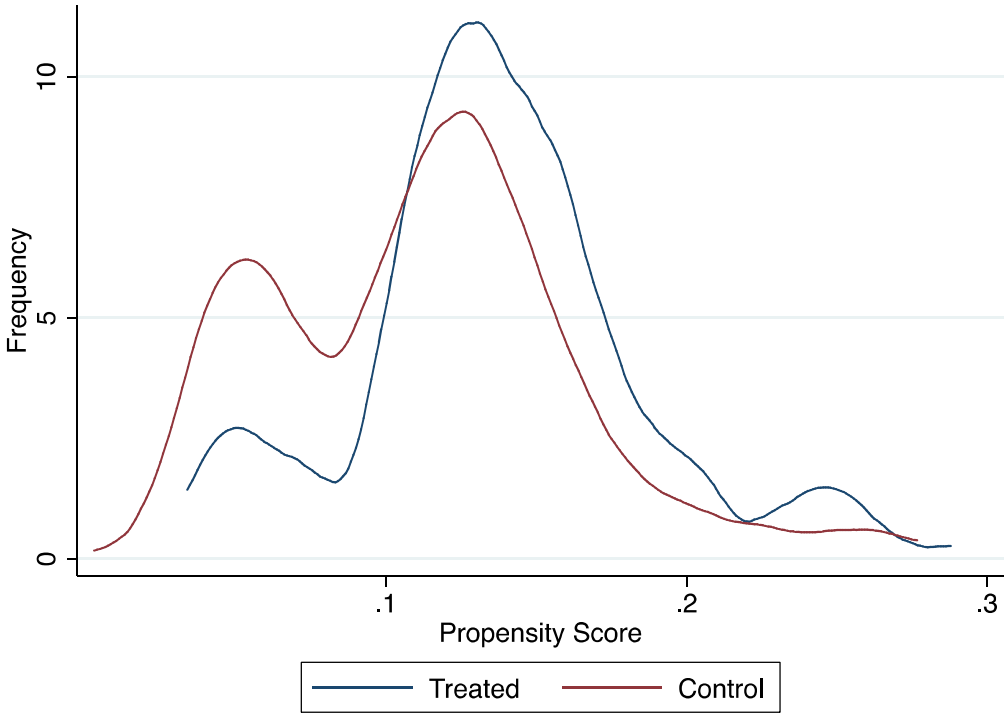
[27] See Abadie et al. (2004).

[28] The report provides an ever more concerning assertion for the nearest neighbor matching in the footnote to Table 4 – i.e., that "a corresponding match could not be found due to differences in socioeconomic or other characteristics" (p. 168). Again, the report does not provide the evidence to validate this claim as Table 4 focuses on program effects and not the similarity of the treatment and potential control villages post-matching.

[29] Note that this is the basis for the first-stage of IPW, the simplest propensity score matching approach (discussed earlier in the section). This propensity score analysis is conducted with only a subset of the variables included in the current MISTI impact evaluation. The variables are included in a note below the figure. We do not advocate the use of IPW for actual estimation, as other augmented approaches such as IPT or CBPS have more favorable properties.

[30] e.g., Becker and Ichino, 2002.

Figure 3: Comparing Treated and Control Villages



Note: Propensity scores are calculated using the following MISTI-provided variables: population, elevation, Pashtun-speaking village, Dari-speaking village, distance to district center, and historical number of NSP projects.

MISTI Remedy and Remaining Issues

Based on the analysis above, the problem of identifying appropriate control villages is perhaps not as severe as the MISTI evaluation team informally suggested. Our simple exercise suggests that the treatment and control villages are sufficiently similar to use for use in the impact evaluation.

Still, we would encourage the researchers to do two things. First, move away from using only the "exact matching" methods used in the current draft. Though these methods have the attractive "exact match" feature, propensity score methods are also relatively intuitive and allow more flexibility in this complex environment as exact matches are not required.[31] The different approaches can be compared as a test of robustness of the findings.[32]

Second, the researchers should consider discussing the variables used for these matching methods with the implementing partners (discussed on page 160). The goal of this effort would be identify whether there are other factors that did affect the likelihood of a village receiving a

[31] Also, if the researchers want to maintain the exact matching property, they can consider the Inverse Probability Tilting approach of Graham, Pinto, and Egel (2011).

[32] A response from Jason Lyall on an earlier version of this report indicated that some such comparisons were in fact done, and that results were similar. We suggest then that these findings be noted in the reports.

project programming which are not included in the currently available data but may have been collected by the implementing partner during their own research.

Implications for Impact Evaluations in Conflict/Complex Settings

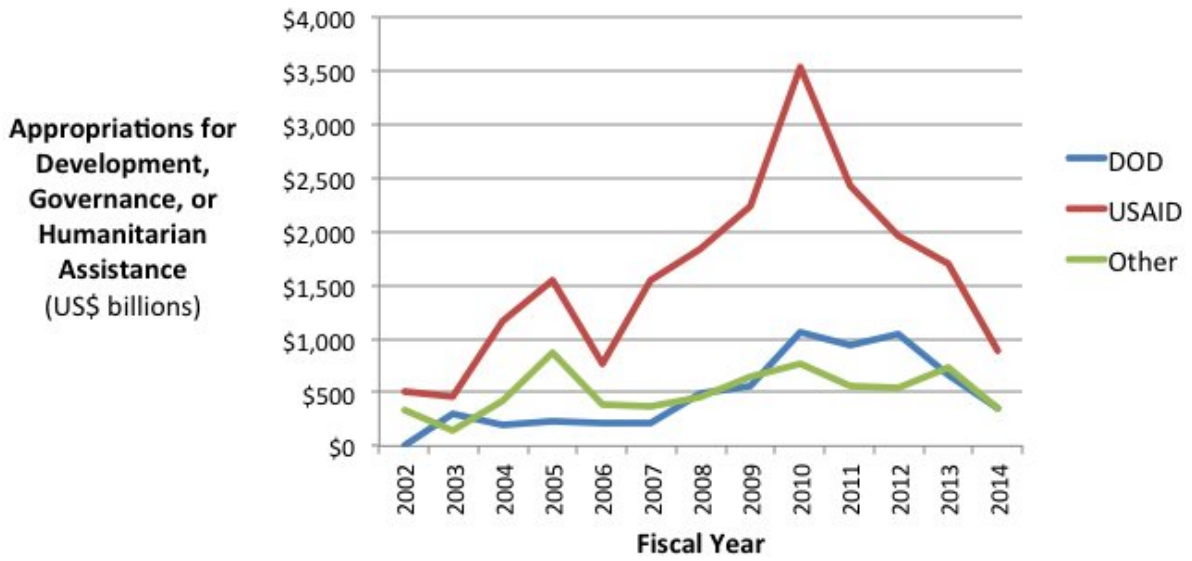
In practice, randomization is often very hard to implement in a conflict setting. Thus, quasi-experimental methods to define a control group, such as matching and difference in difference, are often necessary. However, the reliability of these methods can be greatly enhanced if the researcher coordinates with the implementing partner in two ways. First, the implementing partners should establish clear rules for village eligibility for projects, using measures that are clear and available for all villages. This will ensure that the researcher is able to match on the same variables that go into the determination of program eligibility. Second, the implementing partner and its subcontractors should collect detailed data on site visits to all villages, to provide more information that can be used in the matching. These can go well beyond the indicators used for determining eligibility; in general, the more information the better.

5. Challenge #3: Previous Development Programming

Scope of the Problem

The United States government alone has spent more than $33 billion for development and humanitarian assistance for Afghanistan since 2002; USAID accounted for 62% of these funds (Figure 4). This creates significant challenges for the evaluation of the SIKA and CCI programs.

Figure 4: U.S. Development Spending in Afghanistan



If historical development programming is correlated with the location of SIKA and CCI, this could bias the estimates of the current (SIKA and CCI) program effects.[33] For example a negative correlation between the location of SIKA and CCI programming and historical development programs – i.e., SIKA and CCI were more likely to be implemented in areas that had not previously benefited from development programming – would tend to yield an underestimate of the effect of the current programming; if the earlier programming led to improvements in the control villages in the outcomes of interest (stability and resiliency), then the lack of it in treatment villages would make them appear to have poor outcomes after SIKA or CCI relative to controls even if the current programs were highly successful.

[33] Regardless of the direction of this correlation, the absence of detailed village-level data on historical development programming would lead to omitted-variable bias which would lead to biased point estimates if not properly corrected for (e.g., Heckman and Robb 1985, Esther and Duflo 200?).

The Challenge in the MISTI Context

The impact evaluation addresses this issue in two ways. First, the team has collected relatively detailed data on historical USAID programming conducted as part of the National Solidarity Program (NSP) for inclusion in the analysis. This will allow them to include information on the different types, and number, of NSP projects implemented in their analysis.

However, these data face two central problems. The first is that the NSP accounts for only about $1 billion of all development spending in Afghanistan.[34] Thus, these additional data only capture a small percentage of historical development programming in Afghanistan. The second is that these data only report whether a project was conducted, and not either the success of the project or whether it may have faced challenges in implementation.

The second way in which the evaluation addresses this issue is through the combination of matching and difference-in-difference approaches (discussed earlier). The addition of difference-in-difference means we are comparing *changes* in stability and resiliency indicators in the matched program and non-program villages, not the levels of these indicators. This may significantly alleviate the problem of the correlation of previous and current programming. In essence, since difference-in-difference controls for different baseline levels of the outcome (since it only considers changes over time), it controls for the influence of the prior programs on the outcomes to the extent that this influence is captured by the baseline values of the outcomes.

However, this may not solve the problem completely. The experience with earlier programs may influence the impact of CCI and SIKA themselves. For example the earlier interventions may have made the recipient villages more receptive to the current ones, increasing their success. If earlier and current program are correlated, then we would have an overestimate of the effects of the current programs on the outcomes relative to what the effects would be for villages without the earlier programming. Conversely, villages may become saturated such that after a given pattern of programming, the marginal benefit of new programs is small; this would lead to an underestimate of the benefits of CCI and SIKA.

The above examples demonstrate how the earlier programs can affect not just the baseline levels of the key outcomes, but also the effectiveness of current programs in changing these outcomes. In econometric terms, there is an interaction of past and current programming. The way to handle this is to include interaction terms of the treatment (receiving SIKA or CCI) with an indicator for prior programming in the village. However, this requires having this information for each village, which the evaluation does not have (or at least, did not use). Lacking these interactions means we still have an omitted variables problem, even in a difference-in-difference framework.

[34] http://www.tolonews.com/en/afghanistan/10656-one-billion-dollar-invested-in-national-solidarity-programme

Remaining Issues and Recommended Remedies for the MISTI IE

In addition to the number of NSP projects in each village included in the current analysis, the

There are several centralized databases of development programming in Afghanistan that the impact evaluation should consider including. The first is the centralized database of development programs maintained by the Afghan Ministry of Rural Reconstruction and Development (MRRD); this database, which begins in 2002, has information on programs being implemented across the country. Second, USAID maintains an internal database of their own programming – Afghan INFO.[35] Additionally, the United States Department of Defense maintains a centralized database of data on CERP programming.

Including these data in impact evaluation faces several challenges. The first is that these databases can often be difficult to obtain for issues of either sensitivity or classification, and that some data on development funding by non-U.S. donors (e.g., DFID) may either be unobtainable or simply not exist.[36] The second is that these centralized databases are often inaccurate or incomplete – e.g., projects that were never implemented are sometimes included, locational information is inconsistent.[37] The third is that reporting requirements have evolved over time, so that these data are often inconsistent over time. The fourth is that, even if detailed data are available, the current challenge of identifying the locations of where programs were implemented (discussed in "Challenge #1: Identification of USAID Program Villages") persists. And the fifth, and perhaps most important, is that there is not clear way of identifying the likely heterogeneous effects of this earlier programming.

Implications for Impact Evaluations in Conflict/Complex Settings

Randomization of development programs in conflict and complex settings is hard if not impossible. One aspect of non-randomization is that current program allocation may be related to past programming allocations, leading to omitted variable bias if information on the later is not used in the estimation. However, as already discussed, it is feasible to control for different types of observable characteristics, and past programming can be regarded as one kind of characteristic. Stratifying based on the history of development efforts in an area or using these historical variables as controls in regression would help ameliorate the concern about bias, enhance the reliability of the quasi-experimental methods. Further information on prior development programming would serve as more than just controls. This information would also

---

[35] See http://www.usaid.gov/afghanistan/performance-monitoring-plan. Note that these authors had also heard that USAID had subcontracted a firm to produce estimates of total spending at the district level.

[36] MRRD data used to be available on the web, but are no longer. Afghan Info has never been publicly available, and often not even available to other U.S. Government organizations. CERP data is available and unclassified, though access typically requires a DOD affiliation.

[37] As an example, some CERP projects report the location of the closest military base, some the nearest village, and some the actual location of the project.

provide novel and useful insights on how the history of development programming conditions the benefits from new stability- and resiliency-focused programming.

6. Challenge #4: Design of Stability Index

Description of the Problem

Measuring stability is complicated by both conceptual and technical factors. Conceptually, the MISTI evaluation required a definition of stability that would be broadly applicable to the diverse range of USAID stability-focused programming being conducted in Afghanistan. Thus, the PMP describes stability as:

> "Stability may be defined as the prevailing belief in and support for the decisions and actions of local leaders and government that affect the lives of people in a given community. Stability or instability is thus measured primarily through specific perceptions, and stabilization is measurable through improvements in these perceptions. People in stable areas judge physical security, quality of life, economic opportunities, community relations, and local leaders to be satisfactory. They also generally believe that they receive fair treatment from their local government and legal authorities, and find the daily elements of life to be predictable. Stability is most evident when citizens believe that local leadership and government effectively serves their interests. Stability is strengthened by the presence of a vibrant civil society, ensuring that all groups in society—for example, women and minorities—are able to meaningfully participate in the social and political life of the community." (p.1)

This broad and inclusive definition, which mixes factors that contribute to and are affected by stability, becomes very difficult to operationalize.

The impact evaluation faces two additional challenges. The first is that the USAID programs being evaluated are not designed to directly affect stability. Rather, they focus on providing training, infrastructure, or other programs. A further discussion of this challenge is provided in "Challenge #5" section. Finally, this assessment requires the development of a tool for measuring differences in stability over time and space in a highly heterogeneous developing country. This required that the researchers develop measures that are meaningful (and sensitive to changes) across diverse sociocultural and socioeconomic contexts.[38]

The Challenge in the MISTI Context

The MISTI evaluation uses a single stability metric that is a composite of the following: 35 survey questions from the MISTI evaluation survey (75% of the index); a measure of local

[38] An additional challenge is that potential measures of stability are often only meaningful taking into account other local characteristics that are themselves difficult to measure. As an obvious example, measuring the number of violent attacks in an area without adjusting for population can lead to misleading results about the prevalence of violence. However, population data at both the district and sub-district level in Afghanistan are notoriously unreliable – e.g., LandScan and CIESIN often report different values at both the village- and district-level.

control (10%);[39] a measure of accessibility (10%); and a measure of security incidents in that area (5%). A single index was created as a weighted average of these 38 different factors – with the individual weights assigned to each factor (e.g., 10% for the measure of local control vs. 0.5% for the average response to how satisfied individuals were with their financial situation) based on the views of the research team. This single index approach, while desirable from a policy perspective as it facilitates ease of comparison across time and space, faces two major challenges.

The first challenge is that this single index combines factors that are qualitatively different, that is, represent different things. Unlike other frequently used indices, such as asset indices that include ownership of a range of different durable goods, the stability index includes factors as diverse as "corruption" and "presence of armed groups". There seems to be no clear underlying concept of what stability is that would make the inclusion of these diverse factors in a single measure conceptually clear—certainly people would agree that lower levels of both of these factors would be a positive thing, but how they relate to stability is not transparent (e.g., many stable societies have a great deal of corruption and armed groups may lead to more stability if individual groups are able to dominate specific areas and keep the peace). This problem reflects the conceptual complexity in defining a homogenous stability measure that is appropriate for the diverse USAID programming being implemented.[40]

Combining qualitatively different factors can create significant challenges in interpreting the results from analysis of the index. These challenges are brought out by the analysis in Table 5, which reports the pairwise correlations between each of the 35 different survey questions included in the index; questions are grouped according to indicator and all variables are coded so that a value of 1 has a positive interpretation and a value of zero has a negative interpretation ("positive" meaning favorable, hence presumably pro-stability).[41] Importantly, more than 40% of the pairwise correlations have values that are negative or close to zero and fewer than 40% are greater than 0.1. The frequent lack of strong positive associations—and frequent negative ones—between elements of the index demonstrates that they are not measuring the same thing, and consequently, it is not clear what the stability index is measuring. It also mitigates against finding significant impacts of programs on this index.

The second challenge is that the weights used for constructing the index seem to have been selected based on a priori judgments of the team. These may be well informed, but this process ignores widely accepted approaches for combing multiple, potentially different factors, into a single index (e.g., factor analysis, principal components analysis), which essentially allow the data to determine the appropriate weights as well as assess the similarity or dissimilarity of

---

[39] This was assessed based on enumerator observation.

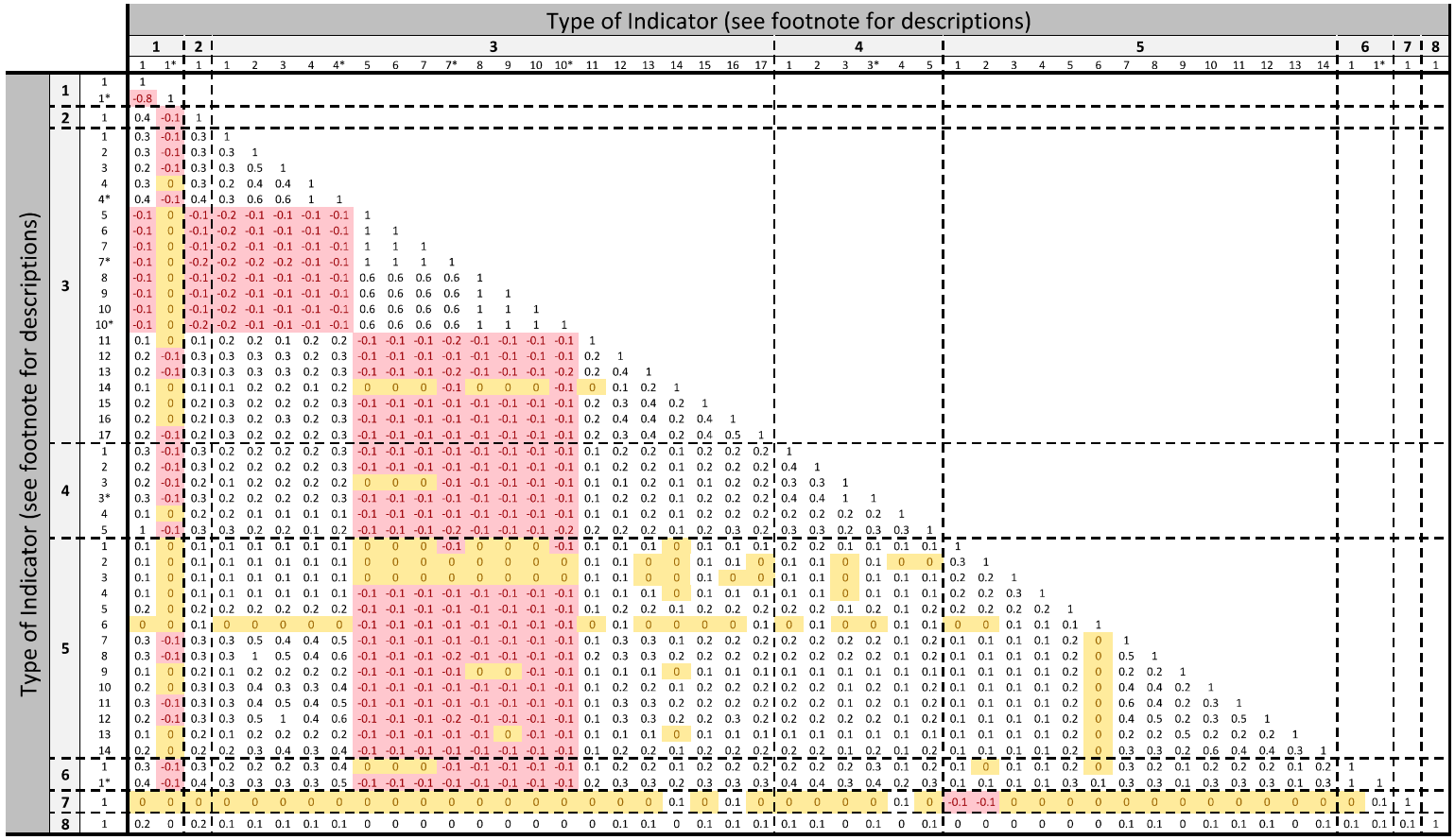
[40] This challenge is closely related to the "Challenge 5: Theory of Change" discussion.

[41] The one exception are the variables with a star next to the numbers. Those variables had a neutral response as a possible categorical response; those neutral responses were coded as missing as indicated in the table notes.

different variables.[42] Note that many of these methods would allow the researcher to identify whether there are multiple "underlying factors" that are driving these observable variables.

[42] Bohrenstedt (2010).

Table 5: Correlation Matrix of Survey Components of Security Index



Notes: All correlations are significant at the 1% level. Pairs of variables in which the second has an asterisk (e.g., "1" and "1*") indicate that the variable had a neutral possible response. In all cases, the row or column with the star is the one in which the starred variable is coded as missing (as is done by the MISTI team in constructing their stability index).

Types of Indicators:
1 = Stability of Area
2 = District moving in right direction
3 = Confidence in local government
4 = Quality of life
5 = Resilience of local communities
6 = Access to basic services
7 = Reported corruption
8 = Presence of armed groups

Remaining Issues and Recommended Remedies for the MISTI IE

The current stability index as it stands is problematic measure, whether it is used for tracking changes in stability over time or for measuring the impact of USAID development programming. In terms of trying to track instability, either for the impact evaluation or the district-level trend analysis, the researchers should consider a two-pronged approach. First, starting with the 35 identified household survey questions and the three additional measures, the researchers should use a principal component analysis or similar approach in order to (1) explore the relationship between these variables, with a particular focus on trying to understand how many different relationships that these variables are measuring, and (2) develop defensible data-driven weights.

Second, though the researchers do discuss individual components separately for the trend analysis, the impact evaluation only assesses changes in the index, that is, across all 38 variables simultaneously. The MISTI impact evaluation should first explore the potential impact that the programming had on its intended outcomes. In some cases these outcomes will be a component of the stability index, and in other cases they will not. This issue is discussed further in subsequent sections.

Third, the researchers should explore approaches for validating and verifying the usefulness of their overall measure. This is possible given the panel nature of the data. One approach would be to do an ex-post assessment of how their early assessment of stability (Fall 2012) correlated with subsequent violence, governance challenges, and challenges in implementing development programs in later waves. Additionally, the research team should coordinate with ISAF and other representatives of the Department of Defense to assess the relationship between their own measures of stability and other, independent, measures of stability based on other data. Understanding the relationship between these, and a clear articulation of any discrepancies could help provide additional confidence in both the overall index as well as for sub-indices.

Implications for Impact Evaluations in Conflict/Complex Settings

A single combined stability index is an attractive feature of an impact evaluation for policymakers. However, such an index is likely to be impractical for assessing the impact of the USAID programming in this context. Rather, the impact evaluation should explore the effects of programming on the direct outcomes of the programs (e.g., improved governance, quality of life, access to economic infrastructure).

7. Challenge #5: Measuring Support for Taliban

Scope of the Problem

Measuring support for the Taliban vis-à-vis the Afghan government faces numerous challenges. First, as the MISTI impact evaluation reports clearly delineate, is that asking respondents directly about their views toward or the activities of non-state actors does not deliver reliable results. Indeed, at least one Afghan survey firm reported that asking direct questions about the activity of non-state actors was often met with trepidation and affected responses throughout the rest of the survey.[43]

Second, there is often regional variation in terms of the terminology that people use for describing these different organizations. Most Afghan survey firms work with locally based researchers that are cognizant of these issues, but the use of different terminology across regions can introduce unwanted heterogeneity in terms of responses.

Third, sophisticated methods designed to avoid these issues – such as the endorsement experiment used by MISTI – are often met by significant "signal to noise" challenges. These methods, which rely on a subtle word variation across surveys (e.g., a single word changed in the middle of a sentence), face several sources of measurement error in the context of the surveys used by MISTI. First, given the (1) length of the survey instrument, (2) the low levels of education among Afghans, and (3) the lack of any incentive to stay focused and attentive, many respondents may simply be respond to questions as quickly as possible without any real consideration of the question. Second, unlike political alignments in the U.S. or other contexts, many Afghans likely simply do not have strong opinions about the Taliban as compared to other actors (e.g., Afghan government). They may have strong regional or tribal allegiances, but regional or national organizations are often irrelevant for these populations.

A fourth challenge is that an impact evaluation using these methods requires that the questions, and the type of information that they elicit, remain stable over time. If the topic questions became irrelevant, increasingly politicized, or increasingly associated with a single group, then they will not provide a useful way of tracking changes over time because the question means something different in two waves.

The Challenge in the MISTI Context

The MISTI approach for measuring support for the Taliban– i.e., the endorsement experiment – faces significant problems in overcoming both the third and fourth challenges

[43] Personal correspondence with Eureka personnel.

outlined above. This emerges from our own examination of the data, which yields through three stylized facts:

- **<u>Stylized Fact #1</u>: There is almost no variation in response across the four different Taliban vs. Afghan Government Issue questions.** This is demonstrated in Table 6, which reports the individual-level correlation across the four different issues questions. The left panel of this table reports the pairwise correlation among the questions that elicit a response about the Afghan government and those in the right panel report the analogous pairwise correlations for individuals asked about the Taliban. The near perfect correlation indicates that interviewees reply nearly identically for each of these questions. Thus it follows that either (1) the responses to the questions are not meaningful or (2) peoples' views toward either the Taliban or the Afghan government trump any single issue.

Table 6: Individual-Level Correlation among Endorsement Experiment Questions

		Afghan Government						Taliban			
		#1	#2	#3	#4			#1	#2	#3	#4
Question	#1	1				Question	#1	1			
	#2	0.9937	1				#2	0.9921	1		
	#3	0.9922	0.9914	1			#3	0.9892	0.9896	1	
	#4	0.9903	0.9893	0.9901	1		#4	0.9886	0.9884	0.9877	1

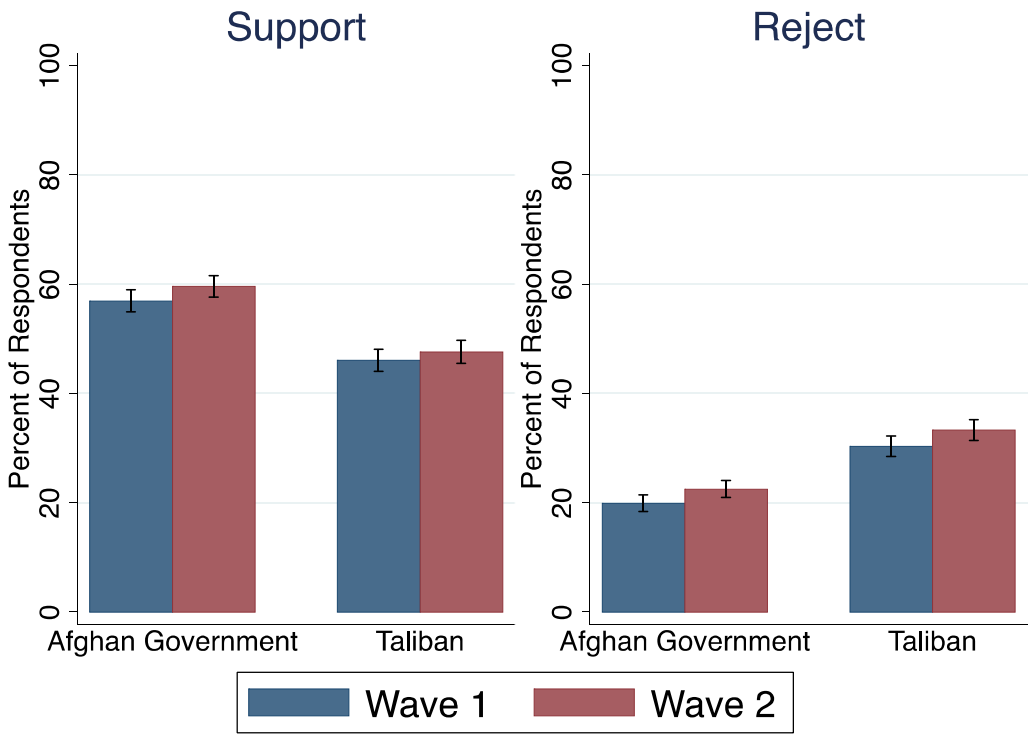
- **<u>Stylized Fact #2</u>: Communities "supportive" of the Afghan government are also more "supportive" of the Taliban.** This is demonstrated in Table 7, which reports the village-level correlation between support for and rejection of either the Afghan government or Taliban. As demonstrated, support for the Afghan government and support for the Taliban are strongly correlated, indicating that villages that are more likely to support the Afghan government are also more likely to support the Taliban. This implausible finding combined with the first stylized fact suggests that these data may only picking up survey effects and not any true effect.

Table 7: Village-Level Correlations among Endorsement Experiment

	Support' Afghan' government	Support' Taliban	Reject' Afghan' government	Reject' Taliban
Support'Afghan'government	1			
Support'Taliban	0.57	1		
Reject'Afghan'government	-0.7476	-0.4703	1	
Reject'Taliban	-0.3489	-0.7684	0.4845	1

- **<u>Stylized Fact #3: The data indicates that support for both the Taliban and the Afghan government has increased from Waves 1 to 3.</u>** This is demonstrated in Figure 5 which reports the share of respondents in the Waves 1 and 3 panel either supporting or rejecting the Taliban or Afghan government – the responses from only one of the four questions is used as an example. In this example, both support and rejection for both groups is increasing across the waves, however three of the four endorsement questions show an increased support for both groups. This also casts doubt on the reliability of the data.

Figure 5: Village-Level Correlations among Endorsement Experiment



Note: Based on responses in the Waves 1 and 3 panels to this question: "Q-51A. It has recently been suggested by the Afghan government [Taliban] that people be allowed to vote in elections to select the members of their district

council. Do you oppose or support such a policy, or are you indifferent to this policy? Do you strongly or only somewhat oppose/support?"

Remaining Issues and Recommended Remedies for the MISTI IE

The above data suggest that this endorsement experiment is unlikely to elicit reliable results for the impact evaluation. We suggest that the MISTI impact evaluation take two steps. The first is that the team attempts to provide additional descriptive data to validate the use of this experimental endorsement approach. However, unlike previous endorsement experiment efforts that included a control group in the experiment (e.g., Lyall et al. 2013), we do not see any clear approach that the impact evaluation team can use for exploring the factors driving the three stylized facts outlined above. Though a control group could be included in subsequent waves, it will be impossible for the team to construct a panel sufficiently long for the impact evaluation. The second, at the very least, is that the limitations identified here could be clearly acknowledged. Ultimately, unless some validation or reasonable interpretation of the findings above is possible, it may be advisable to no longer report results using this method.

Implications for Impact Evaluations in Conflict/Complex Settings

Assessing support for political parties or groups using large-scale national surveys remains an important but difficult to achieve objective. Endorsement experiments are attractive in that they may offer a solution to the problem of respondents' reluctance to state their opinion on these very sensitive topics, but the data presented here illustrates the challenge that they face. Including control groups in these endorsement experiments may help adjust for underlying conditions and augmenting the endorsement experiments with other survey-based approaches designed to explore sympathy for non-state actors (e.g., Cragin 2014) could help.

8. Challenge #6: Theory of Change

Description of Challenge

The intent of the MISTI IE is to assess the impact of USAID-funded "stabilization programs" on stability and resilience. However, properly assessing these programs requires articulating *how* this programming may be influencing these outcomes – a "theory of change".

A clearly articulated theory of change supports both the design of an intervention and its evaluation by providing clear guidance on where and why desired outcomes might be achieved. Without being informed by a theory of change, the evaluation may focus on the wrong outcomes for an intervention, and fail to collect relevant data that may help explain why the intervention works or not.[44]

Additionally, and more fundamentally, without an accepted theory of change directing the design of the programs, USAID implementing partners are likely to implement highly heterogeneous programming. This heterogeneity significantly attenuates the value of any inference from the IE as the researchers will not be able to determine what caused the observed outcomes. We note that heterogeneity per se is not necessarily a problem—it depends on what one is trying to achieve. Infrastructure, training of youth, conflict resolution training, and livelihood support and others may individually be well considered interventions that achieve their direct aims. However, if the programming ultimately is being implemented to achieve stability, it becomes very difficult to understand how this heterogeneous programming is achieving this effect.

A theory of change also provides clarity on when the impact evaluation should be looking for program effects. Are the effects short-term or can the program plausibly achieve sustainable, long-term change? If the goal is to enhance short-term stability in preparation for near-time political change a household survey with a three-month periodicity may be required; however, if the project is an infrastructure project that takes six or nine months to complete, impacts may take considerably longer to detect.

Challenge in the MISTI Context

The need for a well-articulated theory of change seems especially important for the MISTI evaluation since, as stressed earlier, the evaluation is concerned with what are in many cases only indirect impacts of programs that are designed to achieve other specific outcomes (training,

[44] An influential example is the work of Berman, Felter, and Shapiro (2011) who examine develop spending by the military in Iraq. Though they can document a relationship between this spending and attacks against coalition personnel, they do not have the data to provide an empirical explanation of why this relationship was found.

infrastructure, etc.). The links to stability are therefore less immediately obvious and need to be thought through carefully. Unfortunately, the USAID programming in this context—at least from the point of view of the stability objective--does not seem to have been informed by a well-articulated theory of change. For SIKA West, there was not "a defined theory of change articulated in its contract, approved PMP, or work plan".[45] And for CCI there was not a single theory of change; the implementers seem to have followed some combination of both counterinsurgency theory and community cohesion.[46]

This implication for the MISTI IE is that the programs being assessed are likely to be heterogeneous both within and across different implementing partners. This heterogeneity is demonstrated in Table 1 (in the Background section above), which summarizes the different types of projects conducted in the seven different programs. Programming is particularly heterogeneous within IPs; each IP has implemented a mixture of infrastructure, vocational training, and other "community cohesion" exercises. However, the relative share of each of these programs, and the way and locations in which they are implemented, varies significantly across the different IPs.

It goes without saying that each of these programs individually may be beneficial for particular outcomes. However, the MISTI evaluation is concerned with very specific outcomes—stability and resilience. The linkages of the varied programming to these outcomes needs to be clearly articulated in a theory of change, which has not been done. In some cases the links may be weak and understanding this would help interpret the IE findings.

We note additionally that most of these programs were implemented through the local representatives of the relevant ministry, which likely accentuates the heterogeneity in terms of program design, hence effect. In the face of such heterogeneity, estimating an average impact of the interventions, which is what the IE does, may not be very meaningful.

Remaining Issues and Recommended Remedies for the MISTI IE

Assessing the impact of this heterogeneous programming against downstream outcomes, which are themselves poorly defined, is unlikely to give very meaningful results. The MISTI researchers should consider adjusting their evaluation approach in a way that leverages (1) their deep understanding of the geography and intent of each project and (2) broad range of potential outcome measures.

One approach would be to match the specific intent of individual projects to corresponding observable outcomes. As an example, if a project is designed only to improve district government capacity, then the researchers should first test the impact of that particular project against measures that capture perceptions of district governance performance. Impacts on

[45] "Stability in Key Areas – West: Mid-term Performance Evaluation", 26 March 2014.

[46] "Community Cohesion Initiatives: Mid-Term Performance Evaluation Report", 15 May 2014.

stability and resilience, which are more distant outcomes in this case, can of course also be tested. But considering the impacts for which the program was actually designed to influence is important, and will provide a more complete picture of the success or failure of the programs.

The approach could be implemented by estimating a separate quasi-experimental IE regression for the various outcome measures included in the current stability metric. As noted earlier, adjustments would need to be made for multiple hypothesis testing. One approach for implementing this would be a "seemingly unrelated regressions" framework in which multiple estimation equations, one for each outcome variable, are estimated simultaneously.

Implications for Impact Evaluations in Conflict/Complex Settings

Assessing the impact of development programming on downstream outcomes, such as the stability or support for non-state actors, is likely not possible without a clearly articulated theory of change that is mutually understood by both the implementers and the evaluation team. Further, an effective assessment likely requires close oversight of the implementers throughout the lifecycle of the project to understand how programs are being implemented and how local communities are responding to that implementation. Without this, even if significant "positive" results are found, it will not be possible for the evaluator to understand their meaning.

9. Challenge #7: External Validity of MISTI IE Results

Description of the Problem

Internal validity means that the impact evaluation provides reliable estimates of program impact for the individuals or villages studied in the evaluation. Essentially, this happens when the controls are equivalent to the treated group, so that the former provide the appropriate counterfactual of the outcomes of the latter in the absence of the program. External validity in contrast refers to the reliability of the results as a measure of program impact for the target population in general. An evaluation can have a high degree of internal validity (for example, a well-executed randomized controlled trial of a pilot) but poor external validity if (1) the overall target population differs from the individuals or communities in the evaluation, so that the response to the program will differ; or (2) the nature or implementation of the program is different in real world, scaled up conditions compared with a (usually more carefully controlled) pilot evaluation.

Hence for MISTI we need to consider whether the treatment and control villages used in the evaluation are representative of those which are or will be the targets to the USAID, and whether the programming itself is representative of what is and will be carried out generally. If these conditions are not met, the evaluation results will not have external validity.

The Challenge in the MISTI Context

As discussed above, the MISTI team had incomplete information on the specific villages in which USAID programming would be implemented throughout the evaluation. This has two implications for the potential external validity of these results.

The first implication is that MISTI was unable to guarantee a representative sample of the programs being implemented. Unfortunately, as there is no centralized database of where USAID programming was targeted and what conditions were like in those areas, there is no way to assess how representative the programming in the identified villages is of overall USAID programming, hence how serious a threat there is to external validity.

The second implication is that many of the villages surveyed in the baseline Wave 1 became irrelevant by Wave 2, the same was true from Wave 2 to 3, etc. The result, which is demonstrated in Table 8, is that though a total of 4,798 unique villages were surveyed across Waves 1-4, no more than 1,600 of these villages can be included in any two period difference-in-difference analysis and no more than 1,100 can be included in any three period difference-in-difference analysis. As an example, for the impact evaluation presented most recently by which uses Waves 1 and 3, only 980 villages can be included in the analysis.

Table 8: Overlap Across Available Survey Data

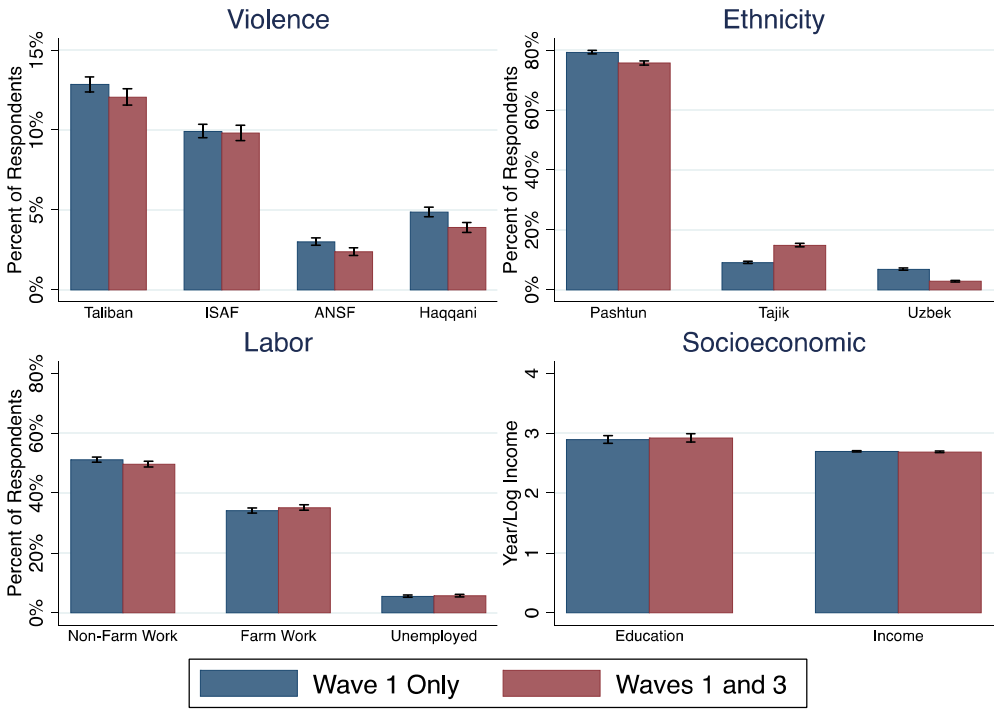
		Waves-of-Data				#-of-Villages
		1	2	3	4	
Pooled		X	X	X	X	4,798
Cross-Sectional-data		X				2,228
			X			2,370
				X		2,548
					X	2,341
Village-Level-Panels	2-Waves	X	X			887
		X		X		980
		X			X	822
			X	X		1,599
			X		X	1,187
	3-Waves	X	X	X		706
		X	X		X	547
		X		X	X	708
			X	X	X	1,106
	4-Waves	X	X	X	X	523

Note: Data using the MISTI village-level data provide slightly different numbers.

The potential implication is that the villages with multiple waves of data may differ from those without these data which must be dropped from the evaluation. Therefore, unlike the baseline data that was originally collected, the data for the evaluation may not be representative of all the villages in the sampled districts that are or will receive programming. This potential loss of representativeness is analyzed in Figures 6 and 7. These figures use two different approaches for comparing the similarity of key socioeconomic and other characteristics among villages that were surveyed in Wave 1 only and those that were surveyed in Waves 1 and 3.

First, Figure 6 compares the means of twelve key characteristics. This approach suggests that there are only limited differences between the villages with data in both periods ("Waves 1 and 3") and those that were not revisited during the third wave ("Wave 1 Only"). Key data on socioeconomic conditions – i.e., employment, education, and income – show no significant differences. Further, experiences with (what is believed to be) Taliban-caused and ISAF-caused casualties are roughly the same. However, there are some limited differences in terms of ethnicity – as Tajik populations seem to be overrepresented and Uzbek populations underrepresented in the panel – and both ANSF- and Haqqani-caused casualties.

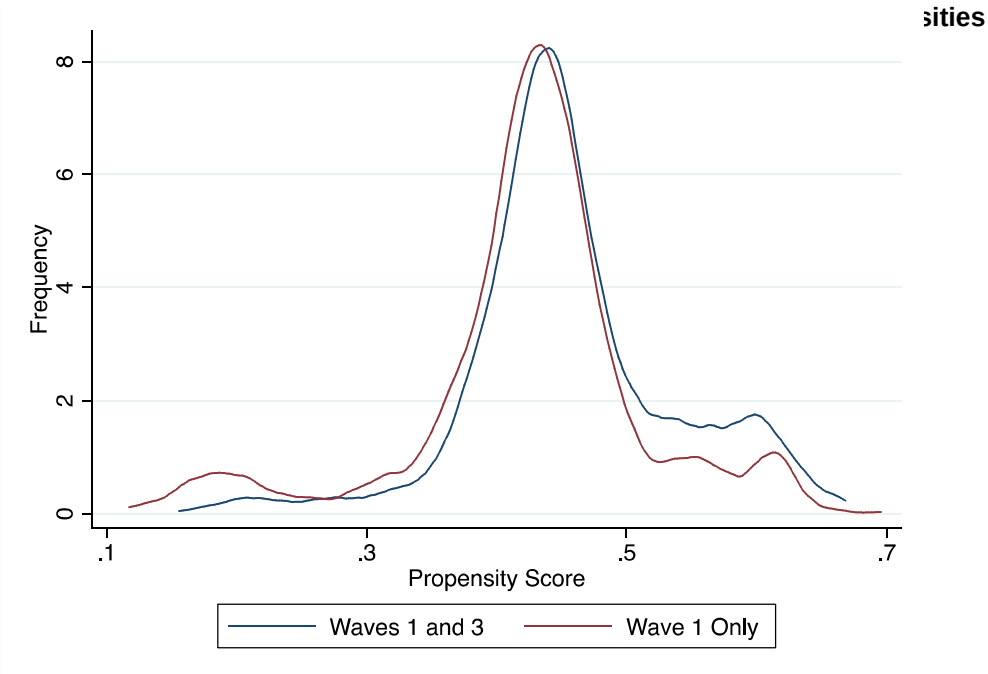
Figure 6: Representativeness of IE Household Panel – Comparing Means



However, while there are not large differences between the means of each group of villages, suggesting that the evaluation sample is broadly representative of all villages in the program districts, Figure 6 reveals that there are still small but significant differences across the two samples. Here we use the same propensity score approach discussed above (Challenge #4), but now we compare the similarity across these two samples (as opposed to comparing treatment and controls).[47] Importantly, not only do these two samples look visually different, the "balancing property" is not satisfied using the standard propensity score approach (e.g., Becker and Ichino, 2002) for the 12 variables from Figure 5. Further, since these two samples differ in the outliers in terms of observables, it is likely that there are significant differences in terms of unobservables as well. Therefore while the comparison of treatment and control villages above (Challenge #2) suggests some degree of internal validity, the external validity of the results is in question.

[47] The "treatment" in this case is being in both Waves 1 and 3 while the "control" is being in Wave 1. Effectively it looks to see whether the 12 variables presented in Figure 9.1 predict significant differences in the likelihood of villages surveyed in Wave being included/excluded in Wave 3.

Figu sities



Remaining Issues and Recommended Remedies for the MISTI IE

As producing IE results that are generalizable to future programming in Afghanistan is likely desirable, the IE should try to address the issues discussed above.

First, in order to address the concern that the treatment villages included in the MISTI are not representative of overall USAID programming, the IE should collect detailed village-level data on every village in which USAID programming was conducted. If sufficient detail on these villages is available, then a matching-type approach (e.g., propensity score) can be used to address this non-random selection. This process can be enhanced significantly if the IE team can coordinate with the IPs to identify the exact criteria that they used for selecting villages (which can be contrasted with the criteria that was used for selection villages included in the household survey).

Second, a similar approach for addressing the non-random attrition from Wave 1 to Wave 3 can be adopted. However, as the "balancing property" could not be satisfied using 12 key demographic variables included in the household survey (discussed above), the IE team should consider a more robust selection of observables.

Implications for Impact Evaluations in Conflict/Complex Settings

In conflict and other complex settings there is always likely to be strong selection in terms of the villages where programs are implemented. This is not a significant challenge in terms of

measuring the impact of the program, but rather in terms of extrapolation of these findings to a broader population.

10. Discussion: Implications for MISTI

This final section reviews the preceding sections to highlight the key lessons learned from our review. The discussion is divided into three parts. The first provides a set of recommendations that the MISTI impact evaluation can implement in producing their Waves 4 and 5 reports, as well as ancillary impact evaluations. These steps can also be applied in a re-analysis of the earlier reports.[48] Note that the MISTI impact evaluation team has already or is planning to implement several of these recommendations. The second part discussed steps that that USAID and MISTI could have undertaken at the outset of the program to ensure a more robust evaluation—and that should be considered for future programming. The third part describes challenges that will always be outside the control of the evaluator, but things that future impact evaluations conducted in conflict areas should be aware of.

Recommendations for Improving MISTI Waves 4 and 5 Reports

1. Assess severity of treatment/control misspecification: At this late stage in the game, project-by-project verification of the location of each project that was either planned or implemented is likely to be infeasible. However, a potential simple and low cost way to assess the severity of misspecification (Challenge #1) is to add a few additional questions to the village-level module to the survey data collection in Wave 5. Specifically, while the survey teams are in the village, the team lead can meet with both village and district officials to determine which villages that implementing partners had visited.[49] This will not allow the impact evaluation team to expand the number of treated villages included in the sample, but it will allow a clear assessment of the severity of misspecification. These data could be combined with MISTI's ongoing treatment village data clarification to allow a reclassification of treatment and control village categorization that can be used to reanalyze the earlier waves' data as well as the data going forward.
2. Conduct power calculations: In order to assess whether the MISTI data have sufficient program villages to measure program effect, the researchers should conduct appropriate power calculations.
3. Move away from "exact matching" quasi-experimental approaches: Given the challenges in the use of quasi-experimental approaches that require on exact matching (highlighted

[48] In several cases, the MISTI evaluation team has, or is, already making efforts to implement these recommendations.

[49] The MISTI evaluation team has suggested that this approach may not yield useful responses at the village-level, as village leaders or officials are unlikely to have visibility on development programming.

by the MISTI team and discussed in the section on Challenge #2), the team should also include other leading quasi-experimental techniques (e.g., propensity score matching approaches) which are more flexible.[50]

4. Work with implementing partners to identify how villages were selected for program participation. The effectiveness of quasi-experimental methods – which essentially involve choosing appropriate, i.e., similar, controls for the treated units – can be enhanced by a clear understanding of how and why villages were selected for participation in one of the USAID programming. Indeed, the implementing partners should have collected data as part of their "sources of instability" assessment for each of these villages. These discussions and assessments would both provide material for the MISTI impact evaluation team to justify their empirical specification for the quasi-experimental methods. Additionally, the research team may learn of other important factors that should be included in the matching; given the richness of the household survey data collected, it should be possible to augment their existing approach with most if not all of these additional factors.[51]
5. Include expanded data on development programming: Both MISTI and USAID should make use of all available data on historical development programming in the areas used for the analysis. This should include MRRD, which is publicly available and includes data on both the NSP program (which is already included in the analysis) and other development programming; unclassified CERP data, which USAID should be able to request from the Department of Defense; as well as the diverse range of USAID collected data on program implementation. These data should be used both to improve the matching of treatment and control villages and in the actual estimation of impacts as control variables; indeed, as discussed in Challenge #3, the evaluation team should explore the interaction of previous programming with current programming as a possible mitigating effect. Further, in addition to simply including the number of projects implemented in an area, the research team should include additional project-specific variables such as total resources allocated (e.g., Berman et al. 2011).
6. Use data-driven methods for stability index. The need to use some kind of stability index is understandable given the objectives of the evaluation. However, the approach currently used is problematic and yields an index that is hard to interpret. The impact evaluation team should use a principal components analysis or factor analysis for developing a more meaningful measure of overall stability. The effectiveness of this

[50] The MISTI impact evaluation team has indicated in private correspondence that they have implemented some propensity score methods; however, the results are not presented in available reports.

[51] The MISTI impact evaluation team has indicated that they have made significant effort to engage with implementing partners; however, they indicated that the implementing partners have been unable to produce any documented record of the results from the "sources of instability" assessment.

method could be enhanced by combining the MISTI data with data from ANQAR, Binna, and other sources.[52]

7. Analyze individual components of stability separately: Given the concerns about the disparate elements of the stability index at least as currently constituted, rather than focusing on a single stability index for the impact evaluation, the researchers should consider various components separately in a seemingly unrelated regressions (or similar approach) framework. Since there are so many individual components (38) grouping them into subindices along clear subject lines is recommended.
8. Validate the stability measure using data from 2012-2013: Given the historically available data, the evaluation team should conduct validation exercises using historical polling data and violence data. If the index does indeed measure stability, then the likelihood of violence should be higher in areas that were judged to be less stable using the index. This should be done for the current index, as well as revised indices developed using the suggestions in recommendation #6 above.
9. Coordinate with ISAF and other representatives to validate stability metric: Given the diversity of U.S. Government organizations involved in assessing instability, the MISTI team should make a robust effort to engage with relevant ISAF officials (e.g., AAG) in order to discuss, validate, and have their approach reviewed. Assessing the relationship between the MISTI and ISAF data, and identifying any discrepancies, could help provide additional confidence in both the overall index as well as the 38 independence sub-indices.
10. Match projects to intended outcomes: Rather than focus on only the reduced form outcomes currently considered—i.e., from program inputs to stability-- the analysis should also validate whether the program is having the intended immediate impact (e.g., improved district governance) as well. Indeed, even if the evidence indicates that USAID programming is enhancing stability, understanding the meaning of those results is impossible without an understanding of whether the programs are achieving their intended development outcomes. Indeed, while the main focus of USAID for this evaluation is on stability, a more logical and sound evaluation approach would be to use the rich survey data for impact evaluations of the programs on the outcomes which they are intended to influence directly, and then proceed to estimating impacts on stability.

[52] E.g., "A Tool for Using Afghan Polling Data to Assess District-Level Changes in Local National Perceptions of Governance, Development, and Security," 2013, PR-432-SOJTF-A (Alex Rothenberg, Miriam Matthews, and Daniel Egel).

Things that Could Have Been Done and Should be Done for Future Evaluations

1. Coordinate with implementing partners from the onset: Close coordination between IPs and evaluation teams is a standard practice now in the impact evaluation literature, and it is of particular importance in developing countries. The main benefit to such a process for MISTI would have been a clearer understanding of where the interventions took place. Given that this coordination typically requires additional human resources cost for the IPs themselves, as they have to attend additional meetings and may need to hire survey specialists to collect or interpret specialized data (e.g., location data) both USAID and MISTI would have to been involved in this type of coordination as well as be willing to support it financially. While MISTI themselves could not have done this, USAID could have modified existing contracts with the implementing partners and provided MISTI-specific resources to the implementing partners to support this effort. It may be noted that while additional resources would be necessary for this, they would likely be a trivial fraction of the total allocated to the evaluation, and have very high returns in terms of the quality of the evaluation.
2. Conduct power calculations: Power calculations conducted in advance of data collection allow the researcher to be judicious in data collection – i.e., to collect sufficient data to detect program effect with expending resources on additional data collection.
3. Stratify based on previous development programming: Given the likelihood that previous development programming could influence the outcomes of the current USAID programming being implemented, USAID should have coordinated to access all available databases on historical development programming and included in the sampling strategy to allow for stratification. Importantly, as the MISTI impact evaluation team indicated that they made substantial efforts to do this at project inception but faced significant difficulties, this process may require USAID-led coordination with other U.S. government foreign government, and international agencies.
4. Cleary articulate theory of change at program commencement. Researchers together with USAID and program developers should articulate a clear theory of change linking the program activities to outcomes and impacts before program commencement. The theory of change would describe how the direct impacts of the programs would lead to improved stability.

Challenges outside MISTI/USAID Control

1. Quality of data on historical development programming: Existing data on development programming does include information on project success, and faces challenges in terms

of accuracy of resources allocated and the specific location of where projects were implemented. Though all these factors will increase the measurement error associated with these data, there is no reasonable way that MISTI could have addressed this.

2. Implementing partner heterogeneity: Implementing programs in highly heterogeneous and conflict-prone settings will often require coordination with a diverse set of implementing partners. While controls can be put in place to try to assess how this affects program implementation and its likely effects, assessing program effects given this heterogeneous treatment can be complicated.
3. External validity of the analysis: Development programming in Afghanistan typically requires targeting the programs to areas that are more receptive to development programs, and successful development programming requires tailoring the programs to local conditions and cultures. Further, it goes without saying that the situation in Afghanistan is unique so that results may not be easily generalized to other conflict settings—though the same can be said of any development related evaluation. As such, while the analysis from MISTI is useful in that it establishes a useful framework for conducting future impact evaluations, the results from this analysis are not necessarily relevant for other potential development programs within Afghanistan or in other conflict-prone contexts.

References

Abadie, A., D. Drukker, J. L. Herr, and G. W. Imbens. 2004. Implementing matching estimators for average treatment effects in Stata. Stata Journal 4(3): 290-311.

Abadie, A. and Imbens, G. (2006) Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**(1): 235–267.

Aigner, D.J. (1973), "Regression with a Binary Independent Variable Subject to Errors of Observation," Journal of Econometrics, 1, 49-60.

Becker, Sascha and Andrea Ichino (2002), "Estimation of Average Treatment Effects Based on Propensity Scores", The Stata Journal, 2(4), 358-377.

Berman, Eli, Jacob N. Shapiro and Joseph H. Felter. "Can Hearts and Minds Be Bought? The Economics of Counterinsurgency in Iraq" Journal of Political Economy, August 2011.

Black, D., S. Sanders, and L. Taylor (2003), "Measurement of Higher Education in the Census and Current Population Survey", Journal of the American Statistical Association, 98, 545-554.

Blackwell, Matthew & Stefano Iacus & Gary King & Giuseppe Porro, 2009. "cem: Coarsened exact matching in Stata," Stata Journal, StataCorp LP, vol. 9(4), pages 524-546, December.

Bohrnstedt, G. (2010). Measurement models for survey research. In P. V. Marsden 8: J. D. Wright (Eds.), Handbook of survey research (2nd ed., pp. 347-404).

Cragin, R. Kim (2014). Resisting Violent Extremism: A Conceptual Model for Non-Radicalization. Terrorism and Political Violence. Vol. 26, Iss. 2, 2014.

Duflo, E. and M. Kremer (2005) "Use of randomization in the evaluation of development effectiveness" in G. Pitman, O. Feinstein, G. Ingram (Eds.), Evaluating Development Effectiveness, Transaction Publishers, New Brunswick, NJ (2005)

Heckman, J. and R. Robb (1985). "Alternative methods for evaluating the impact of interventions: an overview." Journal of Econometrics, 30 (1,2) (1985), pp. 239–267

Frazis, H. and M.A. Loewenstein (2003), "Estimating Linear Regressions with Mismeasured, Possibly Endogenous, Binary Explanatory Variables", Journal of Econometrics, 117, 151-178.

Hirano, K., G. W. Imbens, and G. Ridder (2003, July). Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71 (4), 1161–1189.

Imbens, G. (2004): “Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Survey,” *Review of Economics and Statistics*, 86, 4–30.

Jones, N., Jones, H., Steer, L. and Datta, A. (2009) “Improving Impact Evaluation Production and Use.” Working Paper 300. London: ODI.

Kreider, B. (2010). “Regression coefficient identification decay in the presence of infrequent classification errors”, Review of Economics and Statistics, 92 (4) (2010), pp. 1017–1023.

Jason Lyall, Graeme Blair, and Kosuke Imai, “Measuring Support for Combatants in Wartime: A Survey Experiment in Afghanistan,” American Political Science Review (August 2013).

Millimet, D., 2010. The elephant in the corner: a cautionary tale about measurement error in treatment effects models, Southern Methodist University and IZA Discussion Paper No. 5140.

Ridgeway, G., D. McCaffrey, A. Morral, B.A. Griffin, and L.F. Burgette (2012). twang: Toolkit for weighting and analysis of nonequivalent groups.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89.

USAID (2013), “Technical note in impact evaluation.” September 2013.

200 12th Street South, Suite 1200
Arlington, VA 22202 USA

+1 703 979 7100

MEMORANDUM

October 2, 2014

SUBJECT: MSI-MISTI Comments on the Peer Review of the MISTI Survey and Evaluation Methodology Conducted by the RAND Corporation

TO: USAID/Afghanistan MISTI COR

FROM: MSI Technical Director for MISTI

CC: MISTI Chief-of-Party
MSI Project Manager
The RAND Corporation

Introduction and Background

In March 2014 the USAID/Afghanistan MISTI COR gave MSI the go-ahead to contract an independent third-party peer review of the survey and evaluation methodology utilized by MSI for the MISTI project. In granting the approval, the COR wrote at the time, "This will identify what gaps, if any, need attention and help to ensure the accuracy and strength of our data." After conducting a limited competition for the required work, MSI selected the RAND Corporation (RAND) to conduct the peer review.

Due to the ambitious objectives of the MISTI project and the advanced nature of the methodologies employed by MSI and its MISTI team it was thought that the beneficiaries of the peer review findings would include not only MSI and the USAID Mission in Afghanistan, but the Agency as a whole and the academic and professional evaluation community more broadly.

The peer review was an independent third-party effort which received minimal oversight and guidance from MSI and USAID. Naturally, RAND conducted informational interviews with members of the MISTI team in Kabul and Washington, as well as USAID officials including the COR. MSI was provided by RAND an early draft of the peer review report to review for factual errors-and-omissions – a standard practice in any peer review process.

On September 10, 2014 RAND submitted the final peer review report to the MISTI COR, who in turn requested that MSI-MISTI review and approve it, and (re)transmit to USAID. At the same time, the COR invited the MISTI project to submit brief technical comments on the final report, covering MSI's perspective on the challenges addressed, findings and recommendations discussed in the report. The final, approved deliverable is attached. MSI did not request that RAND make revisions to it, nor were any made since it was sent directly to USAID. The memorandum contains MSI-MISTI comments on the report.

Overall, from MSI's perspective the peer review process was a very positive experience which has provided opportunities for learning and some performance improvement under MISTI. Perhaps more importantly the exercise has resulted in very useful lessons learned for similar subsequent impact evaluation efforts by USAID and others.

Challenges

In addition to the three key findings discussed below, RAND also identified seven technical challenges to the successful implementation of the impact evaluations. Each challenge is addressed below:

1) *Identifying intervention villages:* Treatment villages might be falsely categorized as control villages or control villages might be falsely categorized as treatment villages. Both types of error could create inconclusive and/or invalid findings on impact. We agree with RAND's finding that erroneously categorizing treatment villages as control villages is a much greater challenge than the opposite type of error. If treatments are falsely categorized as controls, then the significance of the impact statistics would be weakened, or the findings would be inconclusive. There is however little danger of a false positive reading. We believe that the coordination required to avoid attribution error is currently in place, and that Survey Waves 4 and 5 will include a large enough number of verified treatment villages to ensure the impact statistics have enough statistical power to yield scientifically credible findings.

2) *Understanding the implementing partners' theories of change*: RAND's description of this challenge suggests that the programs' theories of change may be too vague for evaluating the extent to which the intended change has been achieved. We recognize that, in an ideal world, the theories of change that guide the stabilization programs would provide more specific hypotheses that are more easily testable using conventional methods. The complexity of Afghanistan's environment however requires a more exploratory approach designed to test what types of activities were more or less successful for achieving impact on stability and/or its many component factors.

3) *Lack of comprehensive historical data on development programming*: RAND notes that more data on projects that took place before the current stabilization program might improve MISTI's ability to measure impacts. The utility of such data is however highly speculative. Most of the past programming data that MISTI does not currently have is from CERP and other classified military sources. This data has not been declassified for use by MISTI despite requests made during year one of the program. In any case, CERP projects are unlikely to have significance for the impact evaluation because almost all CERP ended before Survey Wave 1 and its impacts were designed to be short-term (i.e. within a 30 day timeframe). Past programming data from NSP, which MISTI is using for the impact evaluations, is much more important because of the community organization and government linkages that NSP seeks to achieve. Even in the case of NSP, there is only partial overlap with stabilization programming because insecurity led NSP to avoid many districts targeted by SIKA, CCI and KFZ.

4) *Lack of a credible metric for measuring support for the Taliban as compared to the Afghan government:* RAND's discussion of correlations between survey questions reveals a lack of familiarity with the endorsement experiment method used to measure support for the Taliban versus the Karzai Government. The method measures support for several policies, and how presenting each policy as endorsed either by the Taliban or the Karzai Government influences its level of support. Correlations between the individual survey items are immaterial to the endorsement experiment, which involves pooling the answers to the different questions into a single measure of support for Taliban versus Karzai. The metric of support is the difference between levels of support for the same set of policies when the Taliban endorses them compared to an endorsement by Karzai. The endorsement does indeed create a statistically significant difference on levels of support for the same set of policies, which is an indirect method of measuring support for the endorsing entity.

5) *Identifying appropriate control villages:* Control villages are identified through a statistical analysis that identifies a control village that matches each treatment village on all key characteristics, except for the project intervention. Three matching methods are discussed: exact matching, coarsened exact matching, and propensity score matching. The first two methods were used by MISTI in the Wave 3 analysis; RAND used the third as a check on the robustness of the matching performed by MISTI. RAND found that the propensity score matching yielded essentially the same results. Therefore the choice of method used to match villages is a non-issue, MISTI will continue with its chosen coarsened exact matching method given its stronger status as a best practice in field experiment methods, while using propensity score matching for a robustness check.

6) *Developing a defensible metric of stability:* RAND's matrix of correlations between different survey items included in the stability index revealed the opportunity to make changes that will improve the validity of the stability metric and its components. Please refer to our response to Recommendations 6, 8 and 9 below for more details on how we will address this challenge.

7) *External validity of the MISTI evaluation for other stability-focused programming in Afghanistan, or elsewhere in the world:* RAND notes that the villages benefiting from stability programming may not be representative of the overall population, which complicates the effort to generalize the MISTI findings to Afghanistan as a whole and to other countries. The requirement to survey the treatment villages chosen by the stabilization projects may result in a trade-off between the first priority of measuring the impacts of the stabilization projects in fulfillment of the MISTI Task Order, and a secondary consideration of generalizing the findings globally. Indeed, the stabilization programs seem to be intervening in villages where conditions are less stable than the average across Afghanistan, in accordance with their program design. The data certainly provides a wealth of insight for programming in relatively less stable environments in Afghanistan and beyond. We project that Survey Waves 4 and 5 will provide a much richer dataset of treatment villages that may allow for wider generalization to relatively stable areas and villages.

<u>Findings</u>

The Peer Review of the MISTI Survey and Evaluation Methodology prepared by the RAND Corporation included discussion of three key findings:

1) Overall RAND found that MISTI is taking a scientifically credible approach to evaluating direct program impacts. MSI fully concurs with the finding that MISTI provides effective tools for measuring whether programs achieved the impacts for which they were designed. RAND also provided a detailed assessment of the limitations of MISTI's approach, some of which may be overcome by implementing certain recommendations included in the report, as described below.

2) RAND's second major finding included the argument that MISTI may not be able to demonstrate the effect of USAID programming on perceptions of stability. This argument was based on the notion that the stabilization programs provide no "clearly delineated theory of change" for influencing perceptions of stability. In response, we note that the programs are guided by the theory that stability will increase when local sources of instability are first identified, and then addressed by targeted activities. Many different types of activities may be targeted to address one or more sources of instability, depending on the local situation. Perceptions data, such as measures of confidence in the Afghan government, provide direct indications of change caused by the various activities targeted at addressing instability caused by a lack of confidence in government.

3) RAND's third major finding concerns the challenges of coordinating M&E across MISTI, the four SIKA contracts, the two CCI contracts, and the KFZ contract. We believe that coordination challenges are being overcome to the maximum extent possible, given the fact that MISTI cannot impose best practices on other programs because technical direction for each contract is the prerogative of a separate USAID COR. Through mechanisms such as the quarterly community of practice "summits," MISTI has always advocated the adoption of certain technical practices by the different stabilization programs, such as the use of common village lists to record the location of projects, common indicator definitions, and common sub-project status codes and reports. This strenuous and continuous coordination effort has yielded enough success to enable scientifically credible impact evaluations.

Recommendations

RAND provided seven recommendations for addressing the challenges identified above. We discuss each recommendation below:

1) *Assess severity of treatment/control misspecification by augmenting existing MISTI validation effort with a village- or district-level survey module during Wave 5 data collection.*

MISTI will continue to address the issue of treatment/control specification through coordination with the SIKA and CCI implementing partners and other methods, such as validation using aerial imagery, that are described in the RAND report. Attempting to identify treatment villages using a new survey module is highly unlikely to succeed because of the lack of standardized village names and accurate sub-district maps. A new survey module of this type would add little value at best, and be counterproductive at worst. Therefore MISTI does not accept this recommendation.

2) *Conduct power calculations in order to assess whether the MISTI data have sufficient program villages to measure program effect.*

MISTI agrees that statistical power calculations are necessary. Power calculations are typically done prior to programming to ensure that enough treatment/control villages are being selected for the evaluation. In the case of USAID stabilization programming, information on the location and nature of project activities was not known far enough in advance to enable power calculations. Nevertheless, enough treatment and control villages have been sampled over the survey waves to achieve the statistical power required for scientifically credible estimates of stabilization impacts. In Waves 4 and 5 power calculations will be completed to investigate whether valid estimates of treatment effects can be calculated at the level of regional SIKA and CCI projects, as well as at district and provincial levels.

3) *Use propensity score-based quasi-experimental methods (e.g., IPT, CBPS) in addition to only "exact matching" methods and indicate whether the findings are robust to choice of method.*

MISTI accepts this recommendation on matching treatment and control villages. We will continue to use propensity score matching as a robustness check on our preferred method of coarsened exact matching.

4) *Work with implementing partners to identify how villages were selected for program participation.*

From the beginning MISTI has expended effort towards understanding how the IP's select villages for intervention, but they have never been able to clarify how the process works in a way that can be quantified according to fixed criteria. While MISTI will continue working with the IPs on this point, MISTI will follow the intent of this recommendation by analyzing the available data for commonalities among the treated villages.

5) *Include expanded data and project-specific variables on development programming.*

MISTI continually pursues new sources of relevant data. Military sources have so far refused to provide data from CERP and other programs for use by MISTI because this data is classified. Because nearly all CERP activities ended before the MISTI Survey Wave 1, and CERP was designed to achieve short-term impacts, the risk of omitted variable bias arising from CERP is very low. MISTI's use of data from NSP is much more important for eliminating potential omitted variable bias.

6) *Use data-driven methods for deriving the requisite stability index; 8) Validate the stability measure using data from 2012-2013; 9) Coordinate with ISAF and other representatives to validate the stability metric.*

Recommendations 6, 8, and 9 are closely related and therefore addressed in combination: MISTI has already addressed Recommendation 6 using a statistical method called factor analysis to analyze data from all the survey waves (per Recommendation 8). This data-driven method was used to identify specific survey items, and their relative weights, that have been combined into a revised stability index that more conclusively reflects a common, underlying concept of stability. The factor analysis yields variables from the factor scores that will be tested in the impact analysis. While the process revising the stability index was data driven, it nevertheless reflects and validates the theory of stability that was used to design the original stability index and the survey instrument. Recommendation 9 – validating the new stability index using ISAF data – will not be followed because ISAF refuses to provide the necessary, classified data from its ANQAR survey or other sources. Further, the surveys are likely to be incommensurate because of differences in sampling and questionnaires.

7) *Analyze individual or groups of components of stability separately for the impact evaluation; 10) Rather than focus on only the reduced form outcomes currently considered – i.e., from program inputs to stability – the analysis should also evaluate whether the program is having the intended immediate impact (e.g., improved district governance) as well.*

Recommendations 7 and 10 are closely related and addressed in combination: This recommendation reflects MISTI's past and present practice; the Wave 3 report demonstrated how different sub-components of stability were analyzed separately for the impact evaluation. The revised stability index and its sub components should provide impact measures that are more sensitive to changes caused by program activities compared to the old stability index. In the Wave 4 analysis we will also analyze single survey questions, such as confidence in district government, for changes caused by project activities.